SPARSITY VS. STATISTICAL INDEPENDENCE IN ADAPTIVE SIGNAL REPRESENTATIONS: A CASE STUDY OF THE SPIKE PROCESS *

Bertrand Bénichou[1] and Naoki Saito[2]

[1]*Ecole Nationale Supérieure des Télécommunications, 46, rue Barrault, 75634 Paris cedex 13 France*
[2]*Department of Mathematics, University of California, Davis, CA 95616 USA*

(Received ; Revised)

Abstract. Finding a basis/coordinate system that can efficiently represent an input data stream by viewing them as realizations of a stochastic process is of tremendous importance in many fields including data compression and computational neuroscience. Two popular measures of such efficiency of a basis are sparsity (measured by the expected ℓ^p norm, $0 < p \leq 1$) and statistical independence (measured by the mutual information). Gaining deeper understanding of their intricate relationship, however, remains elusive. Therefore, we chose to study a simple synthetic stochastic process called the spike process, which puts a unit impulse at a random location in an n-dimensional vector for each realization. For this process, we obtained the following results: 1) The standard basis is the best both in terms of sparsity and statistical independence if $n \geq 5$ and the search of basis is restricted within all possible orthonormal bases in $\mathbf{R}^n$; 2) If we extend our basis search in all possible invertible linear transformations in $\mathbf{R}^n$, then the best basis in statistical independence differs from the one in sparsity; 3) In either of the above, the best basis in statistical independence is not unique, and there even exist those which make the inputs completely dense; 4) There is no linear invertible transformation that achieves the true statistical independence for $n > 2$.

Key words and phrases: Sparse representation, statistical independence, data compression, basis dictionary, best basis, spike process

1. Introduction

What is a good coordinate system/basis to efficiently represent a given set of images? We view images as realizations of a certain complicated stochastic process whose probability density function (pdf) is not known a priori. *Sparsity* is important here since this is a measure of how well one can compress the data. A coordinate system producing a few large coefficients and many small coefficients has high sparsity for that data. The sparsity of images relative to a coordinate system is often measured by the expected ℓ^p norm of the coefficients where $0 < p \leq 1$. *Statistical independence* is also important since statistically independent coordinates do not interfere with each other (no crosstalk, no error propagation among them). The amount of statistical dependence of input images relative to a coordinate system is often measured by the so-called mutual information, which is a

*This research was partially supported by NSF DMS-99-73032, DMS-99-78321, and ONR YIP N00014-00-1-046.

statistical distance between the true pdf and the product of the one-dimensional marginal pdf's.

Neuroscientists have become interested in efficient representations of images, in particular, images of natural scenes such as trees, rivers, mountains, etc., since our visual system effortlessly reduces the amount of visual input data without losing the essential information contained in them. Therefore, if we can find what type of basis functions are sparsifying the input images or are providing us with the statistically independent representation of the inputs, then that may shed light on the mechanisms of our visual system. Olshausen and Field (1996, 1997) pioneered such studies using computational experiments emphasizing the sparsity. Immediately after their experiments, Bell and Sejnowski (1997), van Hateren and van der Schaaf (1998) conducted similar studies using the statistical independence criterion. Surprisingly, these results suggest that both sparsity and independence criteria tend to produce oriented Gabor-like functions, which are similar to the receptive field profiles of the neurons in our primary visual cortex. However, the relationship between these two criteria has not been understood completely.

These experiments and observations inspired our study in this paper. We wish to deepen our understanding of this intricate relationship. Our goal here, however, is more modest in that we only study the so-called "spike" process, a simple synthetic stochastic process, which puts a unit impulse at a random location in an n-dimensional vector for each realization. It is important to use a simple stochastic process first since we can gain insights and make precise statements in terms of theorems. By these theorems, we now understand what are the precise conditions for the sparsity and statistical independence criteria to select the same basis for the spike process. In fact, we prove the following facts.

- The standard basis is the best both in terms of sparsity and statistical independence if $n \geq 5$ and the search of a basis is restricted within all possible orthonormal bases in $\mathbf{R}^n$.

- If we extend our basis search in all possible invertible linear transformations in $\mathbf{R}^n$, then the best basis in statistical independence differs from the standard basis, which is the best in sparsity.

- In either of the above, the best basis in statistical independence is not unique, and there even exist those which make the inputs completely dense;

- There is no linear invertible transformation that achieves the true statistical independence for $n > 2$.

These results and observations hopefully lead to deeper understanding of the efficient representations of more complicated stochastic processes such as natural scene images.

More information about other stochastic processes, such as the "ramp" process (another simple yet important stochastic process), can be found in Saito et al. (2000, 2001), which also contain our numerical experiments on natural scene images.

The organization of this paper is as follows. In Section 2, we set our notations and terminology. Then in Section 3, we precisely define how to quantitatively measure the sparsity and statistical dependence of a stochastic process relative to a given basis. Using a very simple example, Section 4 demonstrates that the sparsity and statistical independence are two clearly different concepts. Section 5 presents our main results. We prove these theorems in Section 6 and Appendices. Finally, we discuss the implications and further directions in Section 7.

2. Notations and Terminology

Let us first set our notation and the terminology of basis dictionaries and best bases. Let $\boldsymbol{X} \in \mathbf{R}^n$ be a random vector with some unknown pdf $f_{\boldsymbol{X}}$. Let us assume that the available data $\mathcal{T} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ were independently generated from this probability model. The set $\mathcal{T}$ is often called the training dataset. Let $B = (\boldsymbol{w}_1, \ldots, \boldsymbol{w}_n) \in \mathrm{O}(n)$ (the group of orthonormal transformations in $\mathbf{R}^n$) or $\mathrm{SL}^{\pm}(n, \mathbf{R})$ (the group of invertible volume-preserving transformations in $\mathbf{R}^n$, i.e., their determinants are ± 1). The best-basis paradigm, Coifman and Wickerhauser (1992), Wickerhauser (1994), Saito (2000), is to find a basis B or a subset of basis vectors such that the features (expansion coefficients) $\boldsymbol{Y} = B^{-1}\boldsymbol{X}$ are useful for the problem at hand (e.g., compression, modeling, discrimination, regression, segmentation) in a computationally fast manner. Let $\mathcal{C}(B \,|\, \mathcal{T})$ be a numerical measure of *deficiency* or *cost* of the basis B given the training dataset $\mathcal{T}$ for the problem at hand. For very high-dimensional problems, we often restrict our search within the basis dictionary $\mathcal{D} \subset \mathrm{SL}^{\pm}(n, \mathbf{R})$, such as the orthonormal or biorthogonal wavelet packet dictionaries or local cosine or Fourier dictionaries where we never need to compute the full matrix-vector product or the matrix inverse for analysis and synthesis. Under this setting, $B_\star = \arg\min_{B \in \mathcal{D}} \mathcal{C}(B \,|\, \mathcal{T})$ is called the *best basis* relative to the cost $\mathcal{C}$ and the training dataset $\mathcal{T}$. Section 6.3 reviews the concept of the basis dictionary and the best-basis algorithm in details.

We also note that log in this paper implies $\log_2$, unless stated otherwise.

3. Sparsity vs. Statistical Independence

The concept of sparsity and that of statistical independence are intrinsically different. Sparsity emphasizes the issue of compression directly, whereas statistical independence concerns the relationship among the coordinates. Yet, for certain stochastic processes, these two are intimately related, and often confusing. For example, Olshausen and Field (1996, 1997) emphasized the sparsity as the basis selection criterion, but they also as-

sumed the statistical independence of the coordinates. Bell and Sejnowski (1997) used the statistical independence criterion and obtained the basis functions similar to those of Olshausen and Field. They claimed that they did not impose the sparsity explicitly and such sparsity *emerged* by minimizing the statistical dependence among the coordinates. These motivated us to study these two criteria.

First let us define the measure of sparsity and that of statistical independence in our context.

3.1 *Sparsity*

Sparsity is a key property as a good coordinate system for compression. The true sparsity measure for a given vector $\boldsymbol{x} \in \mathbf{R}^n$ is the so-called ℓ^0 quasi-norm which is defined as

$$\|\boldsymbol{x}\|_0 \triangleq \#\{i \in [1, n] : x_i \neq 0\},$$

i.e., the number of nonzero components in $\boldsymbol{x}$. This measure is, however, very unstable for even small perturbation of the components in a vector. Therefore, a better measure is the ℓ^p norm:

$$\|\boldsymbol{x}\|_p \triangleq \left(\sum_{i=1}^{n} |x_i|^p\right)^{1/p}, \quad 0 < p \leq 1.$$

In fact, this is a quasi-norm for $0 < p < 1$ since this does not satisfy the triangle inequality, but only satisfies weaker conditions: $\|\boldsymbol{x} + \boldsymbol{y}\|_p \leq 2^{-1/p'}(\|\boldsymbol{x}\|_p + \|\boldsymbol{y}\|_p)$ where p' is the conjugate exponent of p; and $\|\boldsymbol{x} + \boldsymbol{y}\|_p^p \leq \|\boldsymbol{x}\|_p^p + \|\boldsymbol{y}\|_p^p$. It is easy to show that $\lim_{p \downarrow 0} \|\boldsymbol{x}\|_p^p = \|\boldsymbol{x}\|_0$. See Day (1940), Donoho (1994, 1998) for the details of the ℓ^p norm properties.

Thus, we can use the expected ℓ^p norm minimization as a criterion to find the best basis for a given stochastic process in terms of sparsity:

$$\mathcal{C}_p(B \,|\, \boldsymbol{X}) = E\|B^{-1}\boldsymbol{X}\|_p^p, \tag{3.1}$$

The sample estimate of this cost given the training dataset $\mathcal{T}$ is

$$\mathcal{C}_p(B \,|\, \mathcal{T}) = \frac{1}{N}\sum_{k=1}^{N} \|\boldsymbol{y}_k\|_p^p = \frac{1}{N}\sum_{k=1}^{N}\sum_{i=1}^{n} |y_{i,k}|^p, \tag{3.2}$$

where $\boldsymbol{y}_k = (y_{1,k}, \ldots, y_{n,k})^T = B^{-1}\boldsymbol{x}_k$ and $\boldsymbol{x}_k$ is the kth sample (or realization) in $\mathcal{T}$. We propose to use the minimization of this cost to select the *best sparsifying basis* (BSB):

$$B_p = B_p(\mathcal{T}, \mathcal{D}) = \arg\min_{B \in \mathcal{D}} \mathcal{C}_p(B \,|\, \mathcal{T}).$$

Remark 1. It should be noted that *the minimization of the ℓ^p norm can also be achieved for each realization.* Without taking averages in (3.2), one can select the BSB $B_p = B_p(\{\boldsymbol{x}_k\}, \mathcal{D})$ for each realization $\boldsymbol{x}_k \in \mathcal{T}$. We can guarantee that

$$\min_{B \in \mathcal{D}} \mathcal{C}_p(B \,|\, \{\boldsymbol{x}_k\}) \leq \min_{B \in \mathcal{D}} \mathcal{C}_p(B \,|\, \mathcal{T}) \leq \max_{B \in \mathcal{D}} \mathcal{C}_p(B \,|\, \{\boldsymbol{x}_k\}).$$

For highly variable or erratic stochastic processes, however, $B_p(\{\boldsymbol{x}_k\}, \mathcal{D})$ may significantly change for each k and we need to store more information of this set of N bases if we want to use them to compress the entire training dataset. Whether we should adapt a basis per realization or on the average is still an open issue. See Saito et al. (2000, 2001) for more details.

3.2 *Statistical Independence*

The statistical independence of the coordinates of $\boldsymbol{Y} \in \mathbf{R}^n$ means

$$f_{\boldsymbol{Y}}(\boldsymbol{y}) = f_{Y_1}(y_1) f_{Y_2}(y_2) \cdots f_{Y_n}(y_n),$$

where $f_{Y_k}(y_k)$ is a one-dimensional marginal pdf. The statistical independence is a key property as a good coordinate system for compression and particularly modeling because: 1) damage of one coordinate does not propagate to the others; and 2) it allows us to model the n-dimensional stochastic process of interest as a set of 1D processes. Of course, in general, it is difficult to find a truly statistically independent coordinate system for a given stochastic process. Such a coordinate system may not even exist for a certain stochastic process. Therefore, we should be satisfied with finding the least-statistically dependent coordinate system within a basis dictionary. Naturally, then, we need to measure the "closeness" of a coordinate system $Y_1, \ldots, Y_n$ to the statistical independence. This can be measured by *mutual information* or relative entropy between the true pdf $f_{\boldsymbol{Y}}$ and the product of its marginal pdf's:

$$I(\boldsymbol{Y}) \triangleq \int f_{\boldsymbol{Y}}(\boldsymbol{y}) \log \frac{f_{\boldsymbol{Y}}(\boldsymbol{y})}{\prod_{i=1}^n f_{Y_i}(y_i)} \, \mathrm{d}\boldsymbol{y} = -H(\boldsymbol{Y}) + \sum_{i=1}^n H(Y_i),$$

where $H(\boldsymbol{Y})$ and $H(Y_i)$ are the differential entropy of $\boldsymbol{Y}$ and Y_i respectively:

$$H(\boldsymbol{Y}) = -\int f_{\boldsymbol{Y}}(\boldsymbol{y}) \, \log f_{\boldsymbol{Y}}(\boldsymbol{y}) \, \mathrm{d}\boldsymbol{y}, \quad H(Y_i) = -\int f_{Y_i}(y_i) \, \log f_{Y_i}(y_i) \, \mathrm{d}y_i.$$

We note that $I(\boldsymbol{Y}) \geq 0$, and $I(\boldsymbol{Y}) = 0$ if and only if the components of $\boldsymbol{Y}$ are mutually independent. See Cover and Thomas (1991) for more details of the mutual information.

Suppose $\boldsymbol{Y} = B^{-1}\boldsymbol{X}$ and $B \in \mathrm{GL}(n, \mathbf{R})$ with $\det(B) = \pm 1$. We denote such a group of matrices by $\mathrm{SL}^{\pm}(n, \mathbf{R})$. Note that the usual $\mathrm{SL}(n, \mathbf{R})$ is a subgroup of $\mathrm{SL}^{\pm}(n, \mathbf{R})$. Then, we have

$$I(\boldsymbol{Y}) = -H(\boldsymbol{Y}) + \sum_{i=1}^n H(Y_i) = -H(\boldsymbol{X}) + \sum_{i=1}^n H(Y_i),$$

since the differential entropy is *invariant* under such a invertible volume-preserving linear transformation, i.e.,

$$H(B^{-1}\boldsymbol{X}) = H(\boldsymbol{X}) + \log |\det(B^{-1})| = H(\boldsymbol{X}),$$

because $|\det(B^{-1})| = 1$. Based on this fact, we proposed the minimization of the following cost function as the criterion to select the so-called *least statistically-dependent basis* (LSDB) in Saito (2001):

$$\mathcal{C}_H(B \,|\, \boldsymbol{X}) = \sum_{i=1}^{n} H\left((B^{-1}\boldsymbol{X})_i\right) = \sum_{i=1}^{n} H(Y_i). \tag{3.3}$$

The sample estimate of this cost given the training dataset $\mathcal{T}$ is

$$\mathcal{C}_H(B \,|\, \mathcal{T}) = -\frac{1}{N}\sum_{k=1}^{N}\sum_{i=1}^{n} \log \widehat{f}_{Y_i}(y_{i,k}),$$

where $\widehat{f}_{Y_i}(y_{i,k})$ is an empirical pdf of the coordinate Y_i, which must be estimated by an algorithm such as the histogram-based estimator with optimal bin-width search of Hall and Morton (1993). Now, we can define the LSDB as

$$B_{LSDB} = B_{LSDB}(\mathcal{T}, \mathcal{D}) = \arg\min_{B\in\mathcal{D}} \mathcal{C}_H(B \,|\, \mathcal{T}). \tag{3.4}$$

We note that the differences between this strategy and the standard independent component analysis (ICA) algorithms are: 1) restriction of the search in the basis dictionary $\mathcal{D}$; and 2) approximation of the coordinate-wise entropy. For more details, we refer the reader to Saito (2001) for the former and Cardoso (1999) for the latter.

Now we describe our analysis of some simple stochastic processes.

4. Two-Dimensional Counterexample

This example clearly demonstrates the difference between the sparsity and the statistical independence criteria. Let us consider a simple process $\boldsymbol{X} = (X_1, X_2)^T$ where X_1 and X_2 are independently and identically distributed as the uniform random variable on the interval $[-1, 1]$. Thus, the realizations of this process are distributed as the right-hand side of Figure 1. Let us consider all possible rotations around the origin as a basis dictionary, i.e., $\mathcal{D} = \mathrm{SO}(2, \mathbf{R}) \subset \mathrm{O}(2)$. Then, the sparsity and independence criteria select completely different bases as shown in Figure 1. Note that the data points under the BSB coordinates (45 degree rotation) concentrate more around the origin than the LSDB coordinates (with no rotation) and this makes the data representation sparser. This example clearly demonstrates that the BSB and the LSDB are different in general. One can also generalize this example to higher dimensions.

5. The Spike Process

An n-dimensional *spike process* simply generates the standard basis vectors $\{\boldsymbol{e}_j\}_{j=1}^n \subset \mathbf{R}^n$ in a random order, where $\boldsymbol{e}_j$ has one at the jth entry and all the other entries are zero. One can view this process as a unit impulse located at a random position between 1 and n as shown in Figure 2.

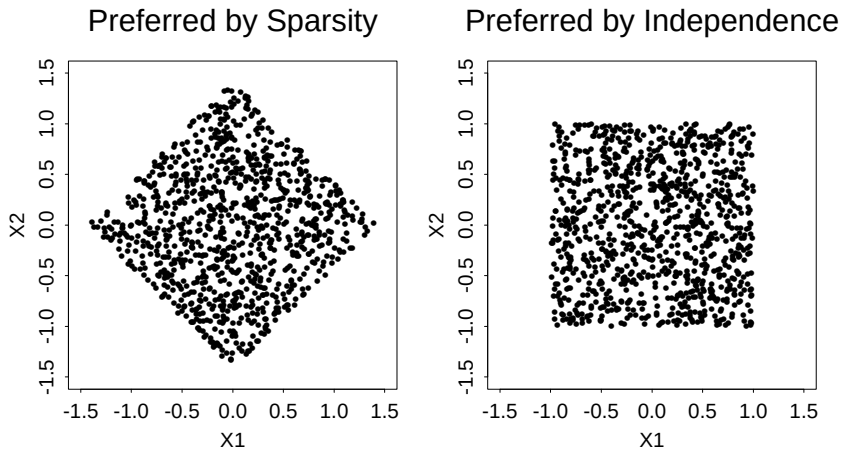

Fig. 1. Sparsity and statistical independence prefer the different coordinates.

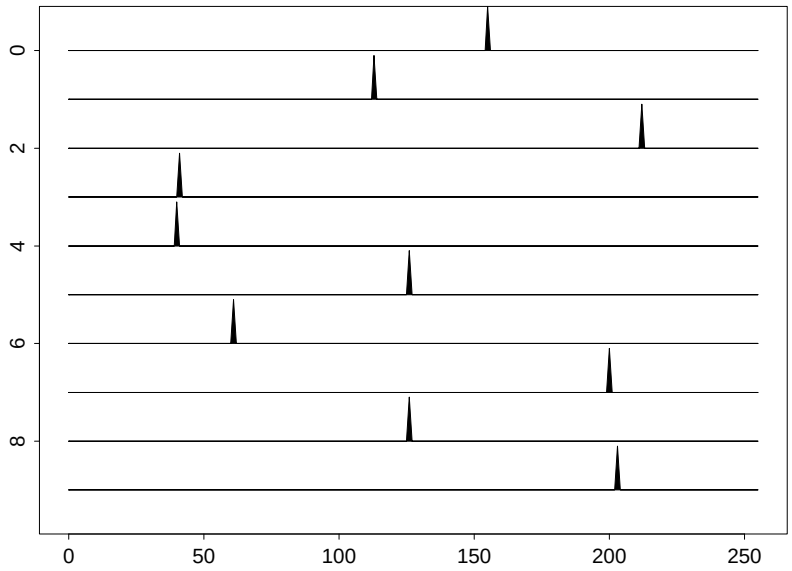

Fig. 2. Ten realizations of the spike process ($n = 256$).

5.1 *The Karhunen-Loève Basis*

Let us first consider the Karhunen-Loève basis of this process from which we can learn a few things.

PROPOSITION 5.1. *The Karhunen-Loève basis for the spike process is any orthonormal basis in* $\mathbf{R}^n$ *containing the "DC" vector* $\mathbf{1}_n = (1, 1, \ldots, 1)^T$.

This means that the KLB is not useful for this process. This is because the spike process is highly non-Gaussian.

5.2 *The Best Sparsifying Basis*

It is obvious that the standard basis is the BSB among $\mathrm{O}(n)$ by construction; an expansion of a realization of this process into any other basis simply increases the number of nonzero coefficients. More precisely, we have the following proposition.

PROPOSITION 5.2. *The BSB for the spike process is the standard basis if* $\mathcal{D} = \mathrm{O}(n)$ *or* $\mathrm{SL}^{\pm}(n, \mathbf{R})$. *If* $\mathcal{D} = \mathrm{GL}(n, \mathbf{R})$, *then it must be a scalar multiple of the identity matrix, i.e.,* aI_n *where* a *is a nonzero constant.*

Remark 2. Note that when we say the basis is a matrix such as aI_n, we really mean that the column vectors of that matrix form the basis. This also means that any permuted and/or sign-flipped (i.e., multiplied by -1) versions of those column vectors also form the basis. Therefore, when we say the basis is a matrix A, we mean not only A but also its permuted and sign-flipped versions of A. This remark also applies to all the propositions, lemmas, and theorems below, unless stated otherwise.

5.3 *Statistical Dependence and Entropy of the Spike Process*

Before considering the LSDB of this process, let us note a few specifics about the spike process. First, although the standard basis is the BSB for this process, it clearly does not provide the statistically independent coordinates. The existence of a single spike at one location prohibits spike generation at other locations. This implies that these coordinates are highly statistically dependent.

Second, we can compute the true entropy $H(\boldsymbol{X})$ for the spike process unlike other complicated stochastic processes. Since the spike process selects one possible vector from the standard basis of $\mathbf{R}^n$ with uniform probability $1/n$, the true entropy $H(\boldsymbol{X})$ is clearly $\log n$. This is one of the rare cases where we know the true high-dimensional entropy of the process.

5.4 *The LSDB among the Haar-Walsh Dictionary*

Our first theorem specifies the LSDB selected from the well-known Haar-Walsh dictionary, a subset of $\mathrm{O}(n)$. This dictionary contains a large number of orthonormal bases (in fact, more than $2^{n/2}$ bases) including the standard basis, the Haar basis (consists of dyadically-scaled and shifted versions of boxcar functions), and the Walsh basis (consisting of square waves). Because the basis vectors in this dictionary are all piecewise constant (except the standard basis vectors), they are often used to analyze and compress discontinuous or blocky signals such as acoustic impedance profiles of subsurface structure. See Wickerhauser (1994), Saito (2000), and Section 6.3 of this paper for the details of this dictionary.

THEOREM 5.1. *Suppose we restrict our search of the bases within the Haar-Walsh dictionary. Then, the LSDB is:*

- *the standard basis if $n > 4$; and*
- *the Walsh basis if $n = 2$ or 4.*

Moreover, the true independence can be achieved only for $n = 2$. Note that n is always a dyadic number in this dictionary.

5.5 *The LSDB among* $\mathrm{O}(n)$

It is curious what happens if we do not restrict ourselves to the Haar-Walsh dictionary. Then, we have the following theorem.

THEOREM 5.2. *The LSDB among* $\mathrm{O}(n)$ *is the following:*

- *for $n \geq 5$, either the standard basis or the basis whose matrix representation is*

$$B_{\mathrm{O}(n)} = \frac{1}{n}\begin{bmatrix} n-2 & -2 & \cdots & -2 & -2 \\ -2 & n-2 & \ddots & & -2 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ -2 & & \ddots & n-2 & -2 \\ -2 & -2 & \cdots & -2 & n-2 \end{bmatrix}; \tag{5.1}$$

- *for $n = 4$, the Walsh basis, i.e.,* $B_{\mathrm{O}(4)} = \frac{1}{2}\begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 \end{bmatrix}$;

- *for* $n = 3$, $B_{\mathrm{O}(3)} = \begin{bmatrix} \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{6}} & \frac{-1}{\sqrt{2}} \\ \frac{1}{\sqrt{3}} & \frac{-2}{\sqrt{6}} & 0 \end{bmatrix}$; *and*
- *for* $n = 2$, $B_{\mathrm{O}(2)} = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$, *and this is the only case where the true independence is achieved.*

Remark 3. There is an important geometric interpretation of (5.1). This matrix can also be written as:

$$I_n - 2\frac{\mathbf{1}_n}{\sqrt{n}}\frac{\mathbf{1}_n^T}{\sqrt{n}}.$$

In other words, this matrix represents the *Householder reflection* with respect to the hyperplane $\{\boldsymbol{y} \in \mathbf{R}^n \mid \sum_{i=0}^n y_i = 0\}$ whose unit normal vector is $\mathbf{1}_n/\sqrt{n}$.

5.6 *The LSDB among* $\mathrm{GL}(n, \mathbf{R})$

Before discussing the LSDB among a larger class of bases, let us remark an important specifics for a discrete stochastic process.

Let $\boldsymbol{X}$ be a random vector obeying a discrete stochastic process with a probability mass function (pmf) $f_{\boldsymbol{X}}$. This means that there are only finite number of possible values (or states) $\boldsymbol{X}$ can take. Clearly the spike process is a discrete process since the only possible values are $\{\boldsymbol{e}_1, \ldots, \boldsymbol{e}_n\}$, the standard basis vectors. Then, for any invertible transformation $B \in \mathrm{GL}(n, \mathbf{R})$ with $\boldsymbol{Y} = B^{-1}\boldsymbol{X}$, be it orthonormal or not, the total entropy of the process before and after the transformation is exactly the same. Indeed, in the definition of discrete Shannon entropy, $-\sum_j p_j \log p_j$, the values that the random variable takes are of no importance; only the number of possible values the random variable can take and its pmf matter. In our case, it is clear that the events $\{\boldsymbol{X} = \boldsymbol{a}_i\}$ and $\{\boldsymbol{Y} = \boldsymbol{b}_i\}$ where $\boldsymbol{b}_i = B^{-1}\boldsymbol{a}_i$ are equivalent; otherwise the transformation would not be invertible. This shows that the corresponding probabilities are equal:

$$\Pr\{\boldsymbol{X} = \boldsymbol{a}_i\} = \Pr\{\boldsymbol{Y} = \boldsymbol{b}_i\}.$$

Therefore, considering the expression of discrete entropy, this proves that

$$H(\boldsymbol{Y}) = H(\boldsymbol{X}),$$

as long as the transformation matrix belongs to $\mathrm{GL}(n, \mathbf{R})$. Note that for the continuous case, this is only true if $B \in \mathrm{SL}^{\pm}(n, \mathbf{R})$. Therefore, for a discrete stochastic process like the spike process, the LSDB among $\mathrm{GL}(n, \mathbf{R})$ can be selected by just minimizing the

sum of the coordinate-wise entropy as (3.4) as if $\mathcal{D} = \mathrm{SL}^{\pm}(n, \mathbf{R})$. In other words, there is no important distinction in the LSDB selection from $\mathrm{GL}(n, \mathbf{R})$ and from $\mathrm{SL}^{\pm}(n, \mathbf{R})$ for discrete stochastic processes. Therefore, we do not have to treat these two cases separately.

THEOREM 5.3. *The LSDB among* $\mathrm{GL}(n, \mathbf{R})$ *with* $n > 2$ *is the following basis pair (for analysis and synthesis respectively):*

$$B^{-1}_{\mathrm{GL}(n,\mathbf{R})} = \begin{bmatrix} a & a & \cdots & \cdots & \cdots & \cdots & a \\ b_2 & c_2 & b_2 & \cdots & \cdots & \cdots & b_2 \\ b_3 & b_3 & c_3 & b_3 & \cdots & \cdots & b_3 \\ \vdots & \vdots & & \ddots & & & \vdots \\ \vdots & \vdots & & & \ddots & & \vdots \\ b_{n-1} & \cdots & \cdots & \cdots & b_{n-1} & c_{n-1} & b_{n-1} \\ b_n & \cdots & \cdots & \cdots & \cdots & b_n & c_n \end{bmatrix}, \tag{5.2}$$

where a, b_k, c_k *are arbitrary real-valued constants satisfying* $a \neq 0$, $b_k \neq c_k$, $k = 2, \ldots, n$.

$$B_{\mathrm{GL}(n,\mathbf{R})} = \begin{bmatrix} \left(1 + \sum_{k=2}^{n} b_k d_k\right)/a & -d_2 & -d_3 & \cdots & -d_n \\ -b_2 d_2/a & d_2 & 0 & \cdots & 0 \\ -b_3 d_3/a & 0 & d_3 & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0 \\ -b_n d_n/a & 0 & \cdots & 0 & d_n \end{bmatrix}, \tag{5.3}$$

where $d_k = 1/(c_k - b_k)$, $k = 2, \ldots, n$.

If we restrict ourselves to $\mathcal{D} = \mathrm{SL}^{\pm}(n, \mathbf{R})$, *then the parameter* a *must satisfy:*

$$a = \pm \prod_{k=2}^{n} (c_k - b_k)^{-1}.$$

Remark 4. The LSDB such as (5.1) and the LSDB pair (5.2), (5.3) provide us with further insight into the difference between sparsity and statistical independence. In the case of (5.1), this is the LSDB, yet does not sparsify the spike process at all. In fact, these coordinates are completely dense, i.e., $\mathcal{C}_0 = n$. We can also show that the sparsity measure $\mathcal{C}_p$ gets worse as $n \to \infty$. More precisely, we have the following proposition.

PROPOSITION 5.3.

$$\lim_{n\to\infty} \mathcal{C}_p\left(B_{\mathrm{O}(n)} \,|\, \boldsymbol{X}\right) = \begin{cases} \infty & \text{if } 0 \leq p < 1; \\ 3 & \text{if } p = 1. \end{cases}$$

It is interesting to note that this LSDB approaches to the standard basis as $n \to \infty$. This also implies that

$$\lim_{n\to\infty} \mathcal{C}_p \left(B_{\mathrm{O}(n)} \,|\, \boldsymbol{X} \right) \neq \mathcal{C}_p \left(\lim_{n\to\infty} B_{\mathrm{O}(n)} \,|\, \boldsymbol{X} \right).$$

As for the analysis LSDB (5.2), the ability to sparsify the spike process depends on the values of b_k and c_k. Since the parameters a, b_k and c_k are arbitrary as long as $a \neq 0$ and $b_k \neq c_k$, let us put $a = 1$, $b_k = 0$, $c_k = 1$, for $k = 2, \ldots, n$. Then we get the following specific LSDB pair:

$$B^{-1}_{\mathrm{GL}(n,\mathbf{R})} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 0 & & & \\ \vdots & & I_{n-1} & \\ 0 & & & \end{bmatrix}, \quad B_{\mathrm{GL}(n,\mathbf{R})} = \begin{bmatrix} 1 & -1 & \cdots & -1 \\ 0 & & & \\ \vdots & & I_{n-1} & \\ 0 & & & \end{bmatrix}.$$

This analysis LSDB provides us with a sparse representation for the spike process (though this is clearly not better than the standard basis). For $\boldsymbol{Y} = B^{-1}_{\mathrm{GL}(n,\mathbf{R})}\boldsymbol{X}$,

$$\mathcal{C}_0 = E\left[\|\boldsymbol{Y}\|_0\right] = \frac{1}{n} \times 1 + \frac{n-1}{n} \times 2 = 2 - \frac{1}{n}.$$

Now, let us take $a = 1$, $b_k = 1$, $c_k = 2$ for $k = 2, \ldots, n$ in (5.2) and (5.3). Then we get

$$B^{-1}_{\mathrm{GL}(n,\mathbf{R})} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 1 & 2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 1 \\ 1 & \cdots & 1 & 2 \end{bmatrix}, \quad B_{\mathrm{GL}(n,\mathbf{R})} = \begin{bmatrix} n & -1 & \cdots & -1 \\ -1 & & & \\ \vdots & & I_{n-1} & \\ -1 & & & \end{bmatrix}. \tag{5.4}$$

The spike process under this analysis basis is completely dense, i.e., $\mathcal{C}_0 = n$. Yet this is still the LSDB.

Finally, from Theorems 5.2 and 5.3, we can prove the following corollary:

Corollary 5.1. *There is no invertible linear transformation providing the statistically independent coordinates for the spike process for $n > 2$. In fact, the mutual information $I\left(B^T_{\mathrm{O}(n)}\boldsymbol{X}\right)$ and $I\left(B^{-1}_{\mathrm{GL}(n,\mathbf{R})}\boldsymbol{X}\right)$ are monotonically increasing as a function of n, and both approaches to $\log \mathrm{e} \approx 1.4427$ as $n \to \infty$.*

Remark 5. Although the spike process is very simple, we have the following interpretation. Consider a stochastic process generating a basis vector randomly at a time selected from some orthonormal basis. Then, both that basis itself is the BSB and the

LSDB among O(n). Theorem 5.2 claims that once we transform the data to the spikes, one cannot do any better than that both in sparsity and independence within O(n). Of course, if one extends the search to nonlinear transformations, then it becomes a different story. We refer the reader to our recent articles Lin et al. (2000, 2001) for the details of a nonlinear algorithm.

6. Proofs of Propositions and Theorems

6.1 *Proof of Proposition 5.1*

PROOF. Let $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)^T$ be a random vector generated by this process. For each of its realizations, a randomly chosen coordinate among these n positions takes the value 1, while the others take the value 0. Hence each X_i, $i = 1, \ldots, n$, takes the values 1 with probability $1/n$ and the value 0 with probability $1 - 1/n$. Let us calculate the covariance of these variables. First, we have:

$$E(X_i) = \frac{1}{n} \times 1 + \left(1 - \frac{1}{n}\right) \times 0 = \frac{1}{n} \quad \text{for } i = 1, \ldots, n$$

$$E(X_i X_j) = \begin{cases} E(X_i^2) = E(X_i) & \text{if } i = j; \\ 0 & \text{if } i \neq j, \end{cases}$$

since one of these two variables will always take the value 0. Let $R = (R_{ij})$ be the covariance matrix of this process. Then, we have:

$$R_{ij} = E(X_i X_j) - E(X_i)E(X_j) = \frac{1}{n}\delta_{ij} - \frac{1}{n^2}$$

We know that a basis is a Karhunen-Loève basis if and only if it is orthonormal and diagonalizes the covariance matrix. Thus, we will now calculate the eigenvalue decomposition of the covariance matrix $R = \frac{1}{n} I_n - \frac{1}{n^2} J_n$, where I_n is the identity matrix of size $n \times n$, and J_n is an $n \times n$ matrix with each entry taking the value 1.

We now need to calculate the determinant:

$$P_R(\lambda) \triangleq \det(\lambda I_n - R) = \begin{vmatrix} \lambda - \frac{1}{n} + \frac{1}{n^2} & \frac{1}{n^2} & \cdots & \frac{1}{n^2} \\ \frac{1}{n^2} & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \frac{1}{n^2} \\ \frac{1}{n^2} & \cdots & \frac{1}{n^2} & \lambda - \frac{1}{n} + \frac{1}{n^2} \end{vmatrix},$$

which is of the generic form:

$$\Delta(a, b) \triangleq \begin{vmatrix} a+b & b & \ldots & b \\ b & a+b & \ddots & \vdots \\ \vdots & \ddots & \ddots & b \\ b & \ldots & b & a+b \end{vmatrix},$$

with the values $a = \lambda - 1/n$ and $b = 1/n^2$. This is calculated by subtracting the last row from all the others, and then adding all $n-1$ columns to the last one.

$$\Delta(a,b) = \begin{vmatrix} a & 0 & \dots & 0 & -a \\ 0 & a & \ddots & \vdots & \vdots \\ \vdots & \ddots & \ddots & 0 & \vdots \\ 0 & \dots & 0 & a & -a \\ b & \dots & \dots & b & a+b \end{vmatrix} = \begin{vmatrix} a & 0 & \dots & 0 & 0 \\ 0 & a & \ddots & \vdots & \vdots \\ \vdots & \ddots & \ddots & 0 & \vdots \\ 0 & \dots & 0 & a & 0 \\ b & \dots & \dots & b & a+nb \end{vmatrix} = a^{n-1}(a+nb). \tag{6.1}$$

Putting $a = \lambda - 1/n$ and $b = 1/n^2$, we have the characteristic polynomial P_R of R as $P_R(\lambda) = \lambda(\lambda - 1/n)^{n-1}$. Hence, the eigenvalues of R are $\lambda = 0$ or $1/n$.

It is now obvious that the vector $\mathbf{1}_n = (1, \dots, 1)^T$ is an eigenvector for R associated with the eigenvalue 0, i.e., $\mathbf{1}_n \in \ker R$. Indeed, we have

$$R\mathbf{1}_n = \left(\frac{1}{n} I_n - \frac{1}{n^2} J_n\right) \mathbf{1}_n = \frac{1}{n}\ \mathbf{1}_n - \frac{1}{n^2}\ n\mathbf{1}_n = 0.$$

Since $\dim \ker R = 1$, $\ker R$ is the one-dimensional subspace spanned by $\mathbf{1}_n$. Considering that R is symmetric and only has two distinct eigenvalues, we know that the eigenspace associated to the eigenvalue $1/n$ is orthogonal to $\ker R$, which is the hyperplane $\{\boldsymbol{y} \in \mathbf{R}^n \mid \sum_{i=1}^n y_i = 0\}$. Therefore, the orthogonal bases that diagonalize R are the bases formed by the adjunction of $\mathbf{1}_n$ to any orthogonal basis of $\ker R^\perp$. The Walsh basis, which consists of oscillating square waves, is such a basis, although it is just one among many. □

6.2 *Proof of Proposition 5.2*

Proof. The case $\mathcal{D} = \mathrm{O}(n)$ is obvious as discussed before this proposition. Therefore, we first prove the case $\mathcal{D} = \mathrm{GL}(n, \mathbf{R})$. To maximize the sparsity, it is clear that the transformation matrix must be diagonal (modulo permutations and sign flips), i.e., $B_p = \mathrm{diag}(a_1, \dots, a_n)$ with $a_k \neq 0$, $k = 1, \dots, n$. The sparsity cost $\mathcal{C}_p$ defined in (3.1) can be computed and bounded in this case as follows:

$$\mathcal{C}_p(B_p \,|\, \boldsymbol{X}) = E\|\boldsymbol{Y}\|_p^p = \frac{1}{n}\sum_{k=1}^n |a_k|^p \geq |a|^p,$$

where $|a| = \min\{|a_1|, \dots, |a_n|\}$. This lower bound is achieved when $B_p = aI_n$, i.e., a nonzero constant times the standard basis. Now, if $\mathcal{D} = \mathrm{SL}^{\pm}(n, \mathbf{R})$, then this constant a must be either 1 or -1 since $\det(B_p) = a^n = \pm 1$ and $a \in \mathbf{R}$. □

6.3 *A Brief Review of the Haar-Walsh Dictionary and the Best-Basis Algorithm*

Before proceeding to the proof of Theorem 5.1, let us first review the Haar-Walsh dictionary and define some necessary quantities.

Let n be a positive dyadic integer, i.e., $n = 2^{n_0}$ for some $n_0 \in \mathbf{N}$. An input vector $\boldsymbol{x} = (x_1, \ldots, x_n)^T$, viewed as a digital signal sampled on a regular grid in time, is first decomposed into low and high frequency bands by the convolution-subsampling operations on the discrete time domain with the pair consisting of a "lowpass" filter $\{h_\ell\}_{\ell=1}^L$ and a "highpass" filter $\{g_\ell\}_{\ell=1}^L$. Let H and G be the convolution-subsampling operators using these filters which are defined as:

$$(H\boldsymbol{x})_k = \sum_{\ell=1}^{L} h_\ell x_{\ell+2(k-1)}, \quad (G\boldsymbol{x})_k = \sum_{\ell=1}^{L} g_\ell x_{\ell+2(k-1)}, \quad k = 1, \ldots, n.$$

We assume the periodic boundary condition on $\boldsymbol{x}$ (whose period is n). Hence, the filtered sequences $H\boldsymbol{x}$ and $G\boldsymbol{x}$ are also periodic with period $n/2$. Their adjoint operations (i.e., upsampling-anticonvolution) H^* and G^* are defined as

$$(H^*\boldsymbol{x})_k = \sum_{1 \le k-2(\ell-1) \le L} h_{k-2(\ell-1)} x_\ell, \quad (G^*\boldsymbol{x})_k = \sum_{1 \le k-2(\ell-1) \le L} g_{k-2(\ell-1)} x_\ell, \quad k = 1, \ldots, 2n.$$

The filter H and G are called conjugate mirror filters (CMF's) if they satisfy the following orthogonality (or perfect reconstruction) conditions:

$$HG^* = GH^* = 0 \quad \text{and} \quad H^*H + G^*G = I,$$

where I is the identity operator. Various design criteria (concerning regularity, symmetry etc.) on the lowpass filter coefficients $\{h_\ell\}$ can be found in Daubechies (1992). The Haar-Walsh dictionary uses the filter pair with the shortest length ($L = 2$) and $h_1 = h_2 = 1/\sqrt{2}$. Once $\{h_\ell\}$ is fixed, the filter G is obtained by setting $g_\ell = (-1)^{\ell-1} h_{L-\ell+1}$. This decomposition process is iterated on both the low and high frequency components. The first level decomposition generates two subsequences $H\boldsymbol{x}$ and $G\boldsymbol{x}$ each of which has length $n/2$. In the case of the Haar-Walsh dictionary, these subsequences are:

$$H\boldsymbol{x} = \left(\frac{x_1 + x_2}{\sqrt{2}}, \ldots, \frac{x_{n-1} + x_n}{\sqrt{2}} \right)^T \quad \text{and} \quad G\boldsymbol{x} = \left(\frac{x_1 - x_2}{\sqrt{2}}, \ldots, \frac{x_{n-1} - x_n}{\sqrt{2}} \right)^T.$$

The second level generates four subsequences, $H^2\boldsymbol{x}$, $GH\boldsymbol{x}$, $HG\boldsymbol{x}$, $G^2\boldsymbol{x}$, each of which is of length $n/4$. If we repeat this process for k times ($k = 0, 1, \ldots, K \le n_0$), then at the kth level, 2^k subsequences $H^k\boldsymbol{x}$, $GH^{k-1}\boldsymbol{x}$, ..., $G^{k-1}H\boldsymbol{x}$, $G^k\boldsymbol{x}$, each of which is of length 2^{n_0-k}, are generated. As a whole there are $(k+1)n$ expansion coefficients (including the original components of $\boldsymbol{x}$). One can iterate this procedure and stop at the Kth level, where $K \le n_0$. These coefficients are naturally organized in the binary tree structure as shown in Figure 3. For future reference, we refer to the tree with $K = n_0$ as the *maximal-depth* tree or the *full* tree. Because of the perfect reconstruction condition on H and G, each decomposition step is also interpreted as a decomposition of the vector space into mutually orthogonal subspaces. Let $\Omega_{0,0}$ denote the n-dimensional Euclidean

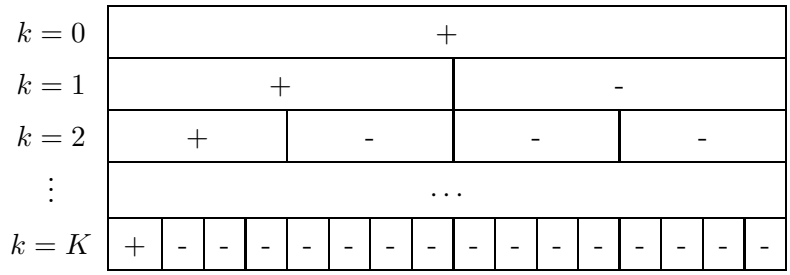

Fig. 3. A table of dictionary coefficients are organized as the binary tree structured table.

space $\mathbf{R}^n$ spanned by the standard basis vectors. Hence, an input vector of length n is an element of $\Omega_{0,0}$. Let $\Omega_{1,0}$ and $\Omega_{1,1}$ be mutually orthogonal subspaces generated by the application of the operators H and G respectively to the parent space $\Omega_{0,0}$. Then, in general, the kth step of the decomposition process ($k = 0, \ldots, K$) can be written as

$$\Omega_{k,\ell} = \Omega_{k+1,2\ell} \oplus \Omega_{k+1,2\ell+1} \quad \ell = 0, \ldots, 2^k - 1.$$

It is clear that $\dim \Omega_{k,\cdot} = 2^{n_0-k}$. For each subspace $\Omega_{k,\ell}$, we associate the basis vectors $\boldsymbol{w}_{k,\ell,m} \in \mathbf{R}^n$, $m = 0, \ldots, 2^{n_0-k} - 1$ which span this subspace. The vector $\boldsymbol{w}_{k,\ell,m}$ is roughly centered at $2^k m$, has length of support $\approx 2^k$, and oscillates $\approx \ell$ times. Note that for $k = 0$, we have the standard basis of $\mathbf{R}^n$. The expansion coefficients computed by the convolution-subsampling operations can be viewed as the inner products between the input vector and these basis vectors although we never need to compute these inner products explicitly. Clearly, we have a redundant set of subspaces in the binary tree. In fact, it is easily proved that there are more than $2^{2^{K-1}}$ possible orthonormal bases in this binary tree; see e.g. Wickerhauser (1994) for the details. Because of this abundance of the bases, such a binary tree of subspaces (or basis vectors) is called a wavelet packet dictionary for general CMF's and *the Haar-Walsh dictionary* if $L = 2$ and $h_1 = h_2 = 1/\sqrt{2}$. Now an important question is how to select the best coordinate system efficiently for the problem at hand from this dictionary.

The "best-basis" algorithm of Coifman and Wickerhauser (1992) first expands an input vector into a specified basis dictionary. Then a complete basis called a *best basis* (BB) which minimizes a certain cost function (such as the sparsity cost $\mathcal{C}_p$ (3.1) or the statistical dependence cost $\mathcal{C}_H$ (3.3); see also Saito (2000) for a variety of cost functions for different problems such as classification and regression) is searched in this binary tree using the divide-and-conquer algorithm. More precisely, let $B_{k,\ell}$ denote a set of basis vectors belonging to the subspace $\Omega_{k,\ell}$ arranged as a matrix

$$B_{k,\ell} = (\boldsymbol{w}_{k,\ell,0}, \ldots, \boldsymbol{w}_{k,\ell,2^{n_0-k}-1}). \tag{6.2}$$

Now let $A_{k,\ell}$ be the best basis for the input signal $\boldsymbol{x}$ restricted to the span of $B_{k,\ell}$ and

let $\mathcal{C}$ be a cost function measuring the deficiency of the nodes (subspaces) such as $\mathcal{C}_p$ or $\mathcal{C}_H$. The following best-basis algorithm "prunes" this binary tree by comparing the cost of each parent node and its two children nodes:
Given an input vector $\boldsymbol{x} \in \mathbf{R}^n$,

Step 0: Choose a basis dictionary $\mathcal{D}$, specify the maximum depth of decomposition K, and an information cost $\mathcal{C}$.

Step 1: Expand $\boldsymbol{x}$ into the dictionary $\mathcal{D}$ and obtain coefficients $\{B_{k,\ell}^T \boldsymbol{x}\}_{0 \le k \le K,\ 0 \le \ell \le 2^k - 1}$.

Step 2: Set $A_{K,\ell} = B_{K,\ell}$ for $\ell = 0, \ldots, 2^K - 1$.

Step 3: Determine the best subspace $A_{k,\ell}$ in the bottom-up manner, i.e., for $k = K - 1, \ldots, 0,\ \ell = 0, \ldots, 2^k - 1$, by

$$A_{k,\ell} = \begin{cases} B_{k,\ell} & \text{if } \mathcal{C}(B_{k,\ell}^T \boldsymbol{x}) \le \mathcal{C}(A_{k+1,2\ell}^T \boldsymbol{x} \cup A_{k+1,2\ell+1}^T \boldsymbol{x}), \\ A_{k+1,2\ell} \oplus A_{k+1,2\ell+1} & \text{otherwise.} \end{cases} \tag{6.3}$$

This algorithm becomes fast if the cost function $\mathcal{C}$ is *additive*, i.e., $\mathcal{C}(\mathbf{0}) = 0$ and $\mathcal{C}(\boldsymbol{x}) = \sum_i \mathcal{C}(x_i)$. Both $\mathcal{C}_p$ of (3.1) and $\mathcal{C}_H$ of (3.3) are clearly additive. If $\mathcal{C}$ is additive, then in (6.3) we have

$$\mathcal{C}(A_{k+1,2\ell}^T \boldsymbol{x} \cup A_{k+1,2\ell+1}^T \boldsymbol{x}) = \mathcal{C}(A_{k+1,2\ell}^T \boldsymbol{x}) + \mathcal{C}(A_{k+1,2\ell+1}^T \boldsymbol{x}).$$

This implies that a simple addition suffices instead of computing the cost of union of the nodes.

Coming back to the Haar-Walsh case, we need a few more definitions for the proof of Theorem 5.1. At each level of the decomposition, the leftmost node (or box) representing the coefficients $H^k \boldsymbol{x}$ is marked by $+$ in Figure 3. This node also corresponds to the subspace $\Omega_{k,0}$. Clearly, each coefficient in this node must be of the form

$$\frac{1}{\sqrt{2^k}} \left(x_{\sigma(1)} + \cdots + x_{\sigma(2^k)} \right) \tag{6.4}$$

where σ is a permutation of $\{1, \ldots, n\}$. We call these nodes and the corresponding coefficients the *positive node* and the *positive coefficients*, respectively. All the other nodes marked by $-$ sign at the kth level corresponding to the subspaces $\Omega_{k,\ell}$, $\ell \neq 0$, contain coefficients of the form:

$$\frac{1}{\sqrt{2^k}} \left(x_{\sigma(1)} + \ldots + x_{\sigma(2^{k-1})} - x_{\sigma(2^{k-1}+1)} - \ldots - x_{\sigma(2^k)} \right). \tag{6.5}$$

These nodes and coefficients are referred to as *negative nodes* and *negative coefficients*, respectively. We note that any descendant node of a negative node must be negative. In fact, only the left child node of a positive node can be positive.

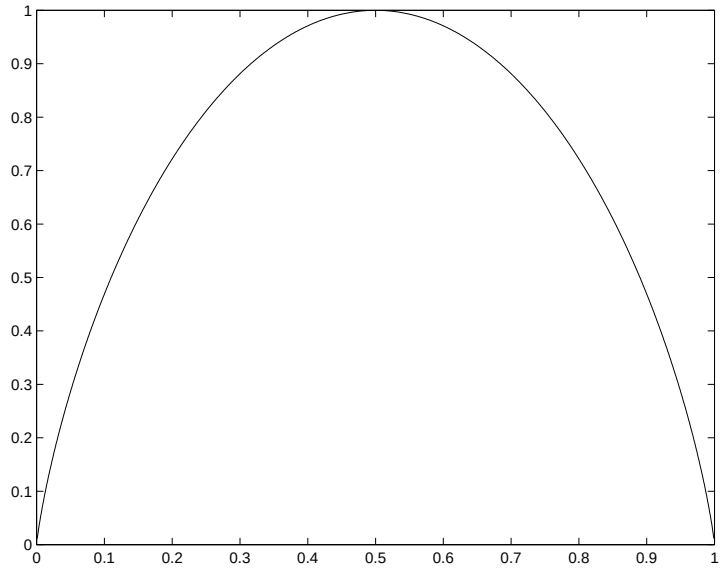

Fig. 4. A plot of $f : x \to -[x \log x + (1-x)\log(1-x)]$.

6.4 *Proof of Theorem 5.1*

PROOF. Let us consider the positive coefficients. The kth-level positive node contains $n/2^k$ coefficients each of which is generated by (6.4), which in the case of the spike process can take only the following values:

- $+1/\sqrt{2^k}$ with probability $2^k/n$;
- 0 with probability $1 - 2^k/n$.

Thus the entropy of each coordinate in the kth-level positive node can be computed as

$$h_+(k) \triangleq -\left(\frac{2^k}{n}\ \log\left(\frac{2^k}{n}\right) + \left(1 - \frac{2^k}{n}\right)\log\left(1 - \frac{2^k}{n}\right)\right) = f\left(\frac{2^k}{n}\right),$$

where

$$f(x) \triangleq -[x\log(x) + (1-x)\log(1-x)], \tag{6.6}$$

which is displayed in Figure 4. The following properties of this function f are basic and will be used repeatedly in this paper:

- For all $x \in [0,1]$, $f(x) \geq 0$ and $f(x) = 0$ if and only if $x = 0$ or $x = 1$;
- For all $x \in [0,1]$, $f(x) = f(1-x)$;
- f is increasing on $[0, 1/2]$, and decreasing on $[1/2, 1]$;
- f is concave on $[0,1]$.

On the other hand, the remaining $n - (n/2^k)$ negative coefficients at level k are computed by (6.5), which can take three different values:

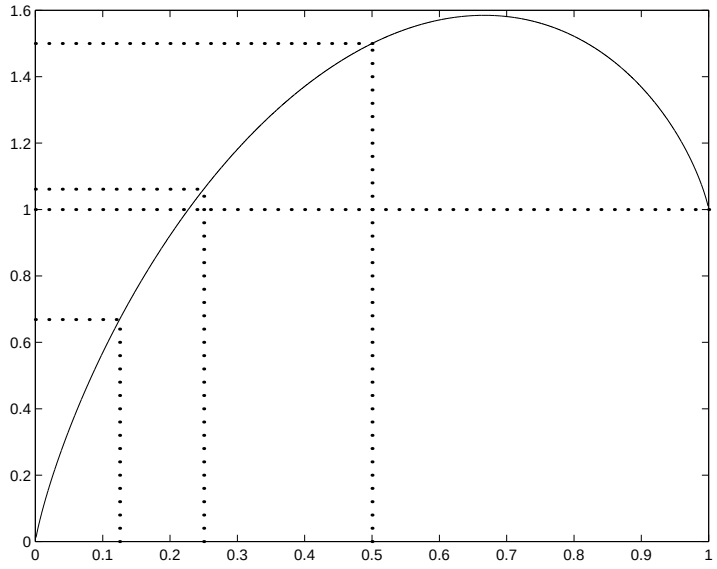

Fig. 5. A plot of $g : x \to -\left[x \log \frac{x}{2} + (1-x)\log(1-x)\right]$.

- $+1/\sqrt{2^k}$ with probability $2^{k-1}/n$;
- $-1/\sqrt{2^k}$ with probability $2^{k-1}/n$;
- 0 with probability $1 - 2^k/n$.

Thus the entropy of each negative coordinate of level k is

$$h_-(k) \triangleq -\left(\frac{2^k}{n}\log\left(\frac{2^{k-1}}{n}\right) + \left(1-\frac{2^k}{n}\right)\log\left(1-\frac{2^k}{n}\right)\right) = g\left(\frac{2^k}{n}\right),$$

where

$$g(x) \triangleq -[x\log(x/2) + (1-x)\log(1-x)] = f(x) + x, \tag{6.7}$$

which is plotted in Figure 5.

The following lemma is used to compare the entropy cost between a parent node and its children nodes of the Haar-Walsh dictionary.

LEMMA 6.1.

$$h_-(k) \le h_-(k+1) \tag{6.8}$$

$$h_+(k) \le \frac{1}{2}\left[h_+(k+1) + h_-(k+1)\right], \tag{6.9}$$

for $k = 1, \ldots, n_0 - 2$.

PROOF. Using the function g defined in (6.7), we have $h_-(k) - h_-(k+1) = g(2^k/n) - g(2^{k+1}/n)$. As shown in Figure 6, the function $g(x) - g(2x)$ is always negative as long as $x = 2^k/n \le 0.43595\cdots$. Since $n = 2^{n_0}$, this implies that $k - n_0 \le \log(0.43595) \approx -1.1977$,

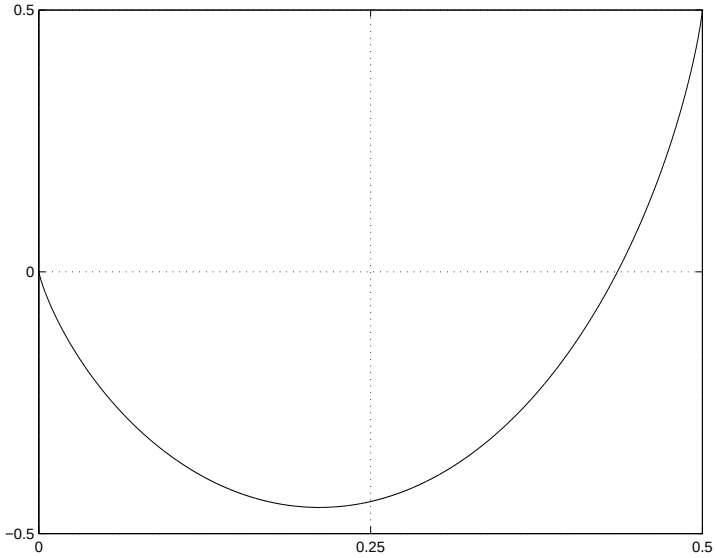

Fig. 6. A plot of the function $g(x) - g(2x)$.

i.e., $k \le n_0 - 2$. Hence we have proved (6.8). To prove (6.9), we have $h_+(k) - \frac{1}{2}[h_+(k+1) + h_-(k+1)] = f(2^k/n) - \frac{1}{2}[f(2^{k+1}/n) + g(2^{k+1}/n)]$. However,

$$\begin{aligned} f(x) - \frac{1}{2}\left[f(2x) + g(2x)\right] &= f(x) - \frac{1}{2}\left[g(2x) - 2x + g(2x)\right] \\ &= (f(x) + x) - g(2x) \\ &= g(x) - g(2x) < 0, \end{aligned}$$

if $x = 2^k/n \le 0.43595$, i.e., $k \le n_0 - 2$ as before. □

Inequality (6.8) implies that the entropy corresponding to a negative coordinate at one level is smaller than that of the level below. Therefore, a negative parent node has smaller entropy than its two negative children nodes provided that the children nodes are *not the bottom leaves*, i.e., if the maximal decomposition level K satisfies $K < n_0$. In fact, we have $h_-(n_0 - 2) \le h_-(n_0 - 1)$, but $h_-(n_0 - 1) \ge h_-(n_0)$. This means that starting from the non-maximal depth negative nodes, the best-basis algorithm always chooses the furthest possible ancestor negative nodes.

As for the positive nodes, from (6.9), we can compare the total entropy of the positive node at level k with that of the two children nodes (positive and negative) as follows:

$$\frac{n}{2^k} h_+(k) \le \frac{n}{2^{k+1}} \left[h_+(k+1) + h_-(k+1)\right],$$

since the parent node contains $n/2^k$ coordinates and each of the children node has $n/2^{k+1}$ coordinates. Therefore, again the parent positive node has smaller entropy than its two children nodes as long as the tree is of non-maximal depth.

These two facts prove that the best-basis algorithm seeking the minimum entropy selects the root node, i.e., the LSDB is the standard basis, if $K \le n_0 - 1$.

Now, we need to consider the case of the maximal-depth tree. Notice that although (6.8) does not hold for $k = n_0 - 1$, the following holds:

$$h_-(n_0 - 3) \leq h_-(n_0),$$

since $g(1) \geq g(1/8)$ (see also Figure 5). This allows us, using the best-basis algorithm, to move up from a pair of bottom leaves not to their immediate parent but to their "great-grandfather", with decreasing entropy, as long as this great-grandfather is still a negative node and $n_0 \geq 3$. We still need to consider what happens if this assumption is false, that is, if we have maximal-depth leaves with positive great-grandfather. The self-similar structure of this tree proves that this problem is equivalent to the general problem with $n_0 = 3$, which we shall now discuss.

$n_0 = 3$ **(i.e., $n = 8$):** Let us show that whatever the set of coordinates chosen among these, the entropy they generate is larger than that of the root node, which is also positive. The entropy of the root node is: $8 \times f(1/8) \simeq 4.34$ bits.

The choice of a basis in this dictionary is equivalent to the choice of a binary tree of depth $K \leq 3$. This reduces to:

- the choice of the level of the positive node in the basis, which also amounts to the choice of the depth of the leftmost leave of the tree.
- the choice of an orthonormal basis of the subspace orthogonal to the chosen positive node.

We note that all the negative coordinates of the tree have larger entropy than those of the bottom leaves: this is derived from $g(1) \leq g(1/4) \leq g(1/2)$ (see Figure 5). Thus, the entropy of any basis with its positive node on level k is larger than $2^{n_0-k} \times f(2^{k-n_0}) + (n - 2^{n_0-k}) \times g(1)$.

Then there are three different cases corresponding to the level of the positive node:

- if the positive node is on the bottom level, then we only have one positive coordinate, and seven negative ones; therefore, the entropy of any such basis is larger than $f(1) + 7 \times g(1) = 7$ bits;
- if the positive node is on level $k = 2$, we have two positive coordinates and six negative ones; thus the entropy of any such basis is larger than $2 \times f(1/2) + 6 \times g(1) \simeq 8$ bits;
- finally, if the positive node is on level $k = 1$, then the entropy of any such basis is larger than $4 \times f(1/4) + 4 \times g(1) \simeq 7.24$ bits.

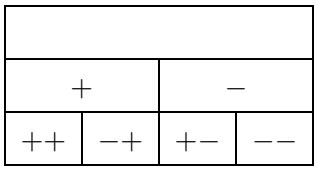

Fig. 7. The Haar-Walsh dictionary table of depth $n_0 = 2$, i.e., $n = 4$.

All these values are larger than the entropy of the root node of this tree, namely the standard basis. Therefore, the standard basis is the LSDB among the Haar-Walsh dictionary for $n_0 = 3$.

$n_0 \geq 3$ **(i.e.,** $n \geq 8$**):** What we saw shows that this is also true for any integer $n_0 \geq 3$ thanks to the self-similar structure of the binary tree dictionary. This ends the proof of the first part of the theorem: the standard basis is the LSDB among the Haar-Walsh dictionary for $n_0 \geq 3$, i.e., $n \geq 8$.

Therefore, we are left to consider the two special cases $n = 2$ and $n = 4$.

$n = 2$**:** In this case, the components of the spike process in the Walsh basis are truly independent. Indeed, the representation of $\boldsymbol{x} = (x_1, x_2)^T$ in the Walsh basis is: $(\frac{x_1+x_2}{\sqrt{2}}, \frac{x_1-x_2}{\sqrt{2}})^T$. The sum of the coordinate-wise entropy of the spike process relative to the Walsh basis is $h_+(1)+h_-(1) = f(1)+g(1) = 0+1 = 1$ bit. That of the standard basis (i.e., the root node) is clearly $2f(1/2) = 2$ bits. Therefore, the Walsh basis always wins over the standard basis. Furthermore, the true entropy of this process is $\log n = \log 2 = 1$ bit, as explained in Subsection 5.3. Therefore, the mutual information of the spike process relative to the Walsh basis is $I(\boldsymbol{Y}) = 1 - 1 = 0$ bit. We therefore have truly independent components for the spike process in this basis for $n = 2$, which is of course the LSDB.

$n = 4$**:** In this case, we consider all possible orthonormal bases in the dictionary exhaustively. Let us mark the table of Figure 7 with + and − signs. We observe:

- each coordinate in the $-+$, $+-$, and $--$ nodes generates the same entropy, $g(1) = 1$ bit;
- the coordinate in the $++$ node generates $f(1) = 0$ bit;
- each coordinate in the $-$ node generates $g(1/2) = 3/2$ bits;
- each coordinate in the $+$ node generates $f(1/2) = 1$ bit.

From these coordinate-wise entropy values, we can compute the entropy of each possible basis in this dictionary as follows:

- the Walsh basis (the level $k = 2$ basis) generates $1 \times 0 + 3 \times 1 = 3$ bits;
- any basis using the + node generates entropy larger than $2 \times 1 + 2 \times 1 = 4$ bits, hence is not the LSDB;
- any basis using the − node generates entropy larger than $2 \times \frac{3}{2} + 1 \times 1 = 4$ bits, hence is not the LSDB;
- the standard basis generates $4 \log 4 - 3 \log 3 \simeq 3.24$ bits, hence is not the LSDB.

Consequently, the LSDB for $n = 4$ is the Walsh basis. This basis does not provide the truly independent components since $I(\boldsymbol{Y}) = 3 - \log n = 3 - \log 4 = 1 \neq 0$.

This concludes the proof of Theorem 5.1. □

6.5 *Coordinate-wise Entropy of the Spike Process*

Before proceeding to the proof of Theorems 5.2 and 5.3, let us consider coordinate-wise entropy of the spike process and define some convenient quantities to characterize a basis in $\mathrm{O}(n)$ or $\mathrm{GL}(n, \mathbf{R})$.

Let us consider an invertible matrix $U = (u_{ij})_{i,j=1,\ldots,n} = B^{-1} \in \mathrm{GL}(n, \mathbf{R})$, and the vector $\boldsymbol{Y} = U\boldsymbol{X}$. Let us consider the ith coordinate of $\boldsymbol{Y}$, $Y_i = \sum_{j=1}^{n} u_{ij} X_j$. For each realization of the spike process $\boldsymbol{X}$, Y_i takes one of the values $\{u_{ij}, j = 1, \ldots, n\}$. More precisely, we have $\Pr\{X_j = 1\} = 1/n$ and $\Pr\{X_j = 0\} = 1 - 1/n$, for $j = 1, \ldots, n$. Thus, if all $\{u_{ij}, j = 1, \ldots, n\}$ were distinct, Y_i would take these values with a uniform pmf. But there is no particular reason that allows us to think $\{u_{ij}, j = 1, \ldots, n\}$ are mutually distinct. Therefore, we shall group these values in "classes" of equality. Let us introduce, for each $i \in \{1, \ldots, n\}$, an integer $k(i)$ equal to the number of distinct values in the ith row vector $\{u_{ij}, j = 1, \ldots, n\}$, and the vector $c(i) = (\alpha_1(i), \ldots, \alpha_{k(i)}(i)) \in \mathbf{N}^{k(i)}$, where each component counts the number of occurrences of each distinct value in the ith row vector. We will call $k(i)$ the *class* of the ith row and $c(i)$ the *index* of that row. Clearly, we have

$$1 \leq k(i) \leq n \quad \text{and} \quad \sum_{\ell=1}^{k(i)} \alpha_\ell(i) = n.$$

For example, with $n = 3$, if we had

$$\begin{cases} Y_1 = X_1 + X_2 + X_3 \\ Y_2 = 5X_1 + 2X_2 + 2X_3 \\ Y_3 = -X_1 + X_2 \end{cases},$$

then we would get

$$\begin{cases} k(1) = 1,\ c(1) = (3) \\ k(2) = 2,\ c(2) = (2, 1) \\ k(3) = 3,\ c(3) = (1, 1, 1) \end{cases},$$

since $\{u_{1j}\} = \{1, 1, 1\}$ in which we find three 1's, $\{u_{2j}\} = \{5, 2, 2\}$ in which we find two 2's, one 5, and $\{u_{3j}\} = \{-1, 1, 0\}$ in which we find one -1, one 1, and one 0.

Let us now examine the coordinate-wise entropy in terms of the quantities we have just defined. Suppose the value u appears $\alpha_\ell(i)$ times in $\{u_{ij}, j = 1, \ldots, n\}$. Then the probability of the event $\{Y_i = u\}$ is $\alpha_\ell(i)/n$. Therefore, we have

$$H(Y_i) = -\sum_{\ell=1}^{k(i)} \frac{\alpha_\ell(i)}{n} \log \frac{\alpha_\ell(i)}{n}.$$

We shall now describe the different values that this coordinate-wise entropy takes as the number of distinct values and their occurrences vary. Because the entropy is a measure of uncertainty, we can intuitively guess that a coordinate with a small class number generates small entropy.

$\underline{k(i) = 1}$**:** This necessarily means that $c(i) = (n)$, i.e., all the $\{u_{ij}, j = 1, \ldots, n\}$ are identical. Since there is no uncertainty about this coordinate, we can expect its entropy to be 0. Indeed, $H(Y_i) = -\sum_{k=1}^{1} \frac{n}{n} \log \frac{n}{n} = 0$.

$\underline{k(i) = 2}$**:** Let us consider the link between the uncertainty and the index $c(i)$. $k(i) = 2$ means that Y_i can take only two distinct values. The least scattered distribution of these two values corresponds to the case $c(i) = (1, n-1)$. This is also the distribution closest to the certain case $k(i) = 1$ and $c(i) = (n)$. We now show that the case $c(i) = (1, n-1)$ generates the smallest entropy. Suppose that Y_i can take two distinct values with index (α_1, α_2), $\alpha_1 + \alpha_2 = n$. In other words, Y_i takes these two values with probability α_1/n and $\alpha_2/n = 1 - \alpha_1/n$, respectively. Without loss of generality, we can assume $\alpha_1 \leq \alpha_2$. Therefore, the entropy of the coordinate Y_i is

$$\begin{aligned} H(Y_i) &= -\left[\frac{\alpha_1}{n} \log \frac{\alpha_1}{n} + \frac{\alpha_2}{n} \log \frac{\alpha_2}{n}\right] \\ &= -\left[\frac{\alpha_1}{n} \log \frac{\alpha_1}{n} + \left(1 - \frac{\alpha_1}{n}\right) \log \left(1 - \frac{\alpha_1}{n}\right)\right] \\ &= f\left(\frac{\alpha_1}{n}\right), \end{aligned}$$

where the function f is defined in (6.6) and shown in Figure 4. Since $\alpha_1 \leq \alpha_2$, it suffices to consider α_1 with $1 \leq \alpha_1 \leq n/2$. So, we have $1/n \leq \alpha_1/n \leq 1/2$, and in this interval, $f(\alpha_1/n)$ is strictly increasing. In other words,

$$f\left(\frac{1}{n}\right) \leq f\left(\frac{\alpha_1}{n}\right) \leq f\left(\frac{1}{2}\right) = 1.$$

Therefore, the entropy is minimal when $\alpha_1 = 1$ and $\alpha_2 = n - 1$. For $\alpha_1 \geq 2$, we clearly have $H(Y_i) \geq f(2/n)$.

$\underline{k(i) \geq 3}$**:** To find a lower bound of $H(Y_i) = -\sum_{\ell=1}^{k(i)} \frac{\alpha_\ell(i)}{n} \log \frac{\alpha_\ell(i)}{n}$, we need the following lemma:

LEMMA 6.2. *Let $k \geq 3$ be an integer, and let $(\alpha_1, \ldots, \alpha_k)$ be a set of strictly positive integers with $\sum_{j=1}^{k} \alpha_j = n$. Then,*

$$\sum_{j=1}^{k} \frac{\alpha_j}{n} \log \frac{\alpha_j}{n} \leq -\left(1 + \frac{2(k-2)}{n}\right) f\left(\frac{1}{n}\right).$$

See Appendix A for the proof of this lemma.

Lemma 6.2 implies that

$$H(Y_i) \geq \left(1 + \frac{2(k-2)}{n}\right) f\left(\frac{1}{n}\right).$$

We can now summarize these results as the following lemma:

LEMMA 6.3. *The coordinate-wise entropy of the spike process after transformed by a basis in* $\mathrm{GL}(n, \mathbf{R})$ *can be computed or bounded as follows:*

(6.10) *if $k(i) = 1$, then $H(Y_i) = 0$;*

(6.11) *if $k(i) = 2$, then* $H(Y_i) \begin{cases} = f(1/n) & \text{if } \alpha_1(i) = 1; \\ \geq f(2/n) & \text{if } 2 \leq \alpha_1(i) \leq n/2; \end{cases}$

(6.12) *if $k(i) \geq 3$, then* $H(Y_i) \geq \left(1 + \frac{2(k-2)}{n}\right) f\left(\frac{1}{n}\right) \geq \left(1 + \frac{2}{n}\right) f\left(\frac{1}{n}\right).$

Let us now come back to our invertible transformation U; we are searching for the LSDB among $\mathrm{O}(n)$ or $\mathrm{GL}(n, \mathbf{R})$. This means that the cost of the LSDB, i.e., the sum of the coordinate-wise entropy of the LSDB coordinates, cannot be larger than that of the standard basis. Therefore we will always keep the standard basis in mind as a reference basis with which we shall compare the performance of all other bases.

The standard basis corresponds to $U = I_n$. Every row of the standard basis has index $k(i) = 2$ and $c(i) = (1, n-1)$. Hence the entropy cost of the standard basis is

$$\mathcal{C}_H(I_n \,|\, \boldsymbol{X}) = n \times f(1/n) = n \log n - (n-1) \log(n-1). \tag{6.13}$$

We saw that, assuming $k(i) > 1$, $H(Y_i) \geq f(1/n)$, with equality if and only if $k(i) = 2$ and $c(i) = (1, n-1)$. Therefore a basis with $k(i) > 1$ for every $i \in \{1, \ldots, n\}$ has no chance to win over the standard basis, and the best thing one can do with such a basis is to match the entropy with that of the standard basis, i.e., a basis with $k(i) = 2$ and $c(i) = (1, n-1)$ for every i.

So, the only chance to beat the standard basis is to have some "class 1" rows (i.e., $k(i) = 1$) in a basis. However, we will never find an invertible matrix with more than one class 1 rows. Indeed, a class 1 row is necessarily proportional to $\mathbf{1}_n^T = (1, 1, \ldots, 1)$, and it is evident that more than one class 1 rows cannot exist in any invertible matrix.

6.6 *Proof of Theorem 5.2*

PROOF. Let us start with a simple remark. If we assume that B is an orthonormal basis, then $U = B^{-1} = B^T$. Hence the rows of U are in fact the basis vectors of this basis. In the case of an orthonormal matrix, the presence of one row of class 1 imposes a constraint on the other rows, since these rows must form an orthonormal basis. The following lemma describes one of these constraints.

LEMMA 6.4. *If $k(1) = 1$, then it is impossible to have two class 2 rows with index $(1, n-1)$ in a matrix $U \in \mathrm{O}(n)$. In other words, If $k(1) = 1$, then there do not exist $i_1, i_2 \in \{1, \ldots, n\}$ such that $i_1 \neq i_2$ and $c(i_1) = c(i_2) = (1, n-1)$.*

The proof of this lemma can be found in Appendix B.

Hence, assuming that $k(1) = 1$, we can have at most one row of class 2 with index $(1, n-1)$. All the other rows will be of either class $k(i) > 2$ or class $k(i) = 2$ with index $(\alpha_1, n-\alpha_1)$, $1 < \alpha_1 \leq n/2$. Considering the minimization of the sum of the coordinate-wise entropy, we must have one row of class 1 and one row of class 2 with index $(1, n-1)$. All the other cases always increase the entropy, i.e., dependency. From (6.11) and (6.12), the entropy of a row with either $k(i) > 2$ or $k(i) = 2$ with index $(\alpha_1, n-\alpha_1)$, $1 < \alpha_1 \leq n/2$ is bounded from below as

$$H(Y_i) \geq \min\left(\left(1+\frac{2}{n}\right) f\left(\frac{1}{n}\right), f\left(\frac{2}{n}\right)\right) = \min\left(\frac{2}{n} f\left(\frac{1}{n}\right), f\left(\frac{2}{n}\right) - f\left(\frac{1}{n}\right)\right) + f\left(\frac{1}{n}\right).$$

Therefore, combining this with (6.10) for $k(1) = 1$ and (6.11) for $\alpha_1 = 1$, we have

$$\sum_{i=1}^{n} H(Y_i) \geq 0 + f\left(\frac{1}{n}\right) + (n-2)\left[\min\left(\frac{2}{n} f\left(\frac{1}{n}\right), f\left(\frac{2}{n}\right) - f\left(\frac{1}{n}\right)\right) + f\left(\frac{1}{n}\right)\right]. \tag{6.14}$$

We now use the following lemma:

LEMMA 6.5. *For $n \geq 6$,*

$$\min\left(\frac{2}{n} f\left(\frac{1}{n}\right), f\left(\frac{2}{n}\right) - f\left(\frac{1}{n}\right)\right) = \frac{2}{n} f\left(\frac{1}{n}\right).$$

PROOF. Let us define a function: $r(x) \triangleq x\left[\frac{2}{x} f\left(\frac{1}{x}\right) - \left(f\left(\frac{2}{x}\right) - f\left(\frac{1}{x}\right)\right)\right]$ for $x \geq 2$, where f is defined in (6.6). This is a continuous and monotonically-decreasing function for $x \geq 2$, since

$$r'(x) = -\frac{2}{x^2}\log(x-1) + \log\frac{x-2}{x-1} < 0 \quad \text{for } x \geq 2.$$

Moreover, we have $r(5) \approx 0.199$ and $r(6) \approx -0.310$, and we can find a zero of $r(x)$ numerically, i.e., $r(x^*) = 0$ where $x^* \approx 5.3623$. These prove that this function is negative if $x \geq x^*$. Therefore, for each integer $n \geq 6$, $r(n) < 0$, i.e.,

$$\frac{2}{n} f\left(\frac{1}{n}\right) < f\left(\frac{2}{n}\right) - f\left(\frac{1}{n}\right).$$

□

Using this lemma for $n \geq 6$, (6.14) can be written as

$$\sum_{i=1}^{n} H(Y_i) \geq f\left(\frac{1}{n}\right) + (n-2)\left[\frac{2}{n} f\left(\frac{1}{n}\right) + f\left(\frac{1}{n}\right)\right] = \left[\frac{2(n-2)}{n} + n - 1\right] f\left(\frac{1}{n}\right).$$

Therefore, if we compare the mutual information of the new coordinates to that of the standard basis, we have

$$I(\boldsymbol{Y}) - I(\boldsymbol{X}) \geq \left[\frac{2(n-2)}{n} + n - 1\right] f\left(\frac{1}{n}\right) - n f\left(\frac{1}{n}\right) = \left[\frac{2(n-2)}{n} - 1\right] f\left(\frac{1}{n}\right),$$

That is,

$$I(\boldsymbol{Y}) - I(\boldsymbol{X}) \geq \frac{n-4}{n} f\left(\frac{1}{n}\right) > 0.$$

Thus, $B = U^{-1} = U^T$ is not the LSDB. We have therefore proved that any orthonormal basis yields a larger mutual information than the standard basis for the spike process for $n \geq 6$.

We can summarize our results so far.

- For $n \geq 6$, the standard basis is the LSDB among $\mathrm{O}(n)$.
- Any basis that yields the same mutual information as the standard basis necessarily consists of only class 2 rows with index $(1, n-1)$.

Now the question is whether there is any other basis except the standard basis satisfying this condition. The following lemma concludes the proof of Theorem 5.2 for $n \geq 6$.

LEMMA 6.6. *For $n \geq 2$, an orthonormal basis consisting of class 2 rows with index $(1, n-1)$ other than the standard basis is uniquely (modulo permutations and sign flips as described in Remark 2) determined as (5.1), i.e.,*

$$B_{\mathrm{O}(n)} = \frac{1}{n} \begin{bmatrix} n-2 & -2 & \cdots & -2 \\ -2 & n-2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & -2 \\ -2 & \cdots & -2 & n-2 \end{bmatrix}.$$

The proof of this lemma can be found in Appendix C. Note that this matrix becomes a permuted and sign-flipped version of I_2 when $n = 2$, and approaches to the identity matrix as $n \to \infty$.

We now prove the particular cases, $n = 3, 4, 5$ in Theorem 5.2. For these small values of n, we cannot use Lemma 6.5 anymore since we have

$$\min\left(\frac{2}{n} f\left(\frac{1}{n}\right), f\left(\frac{2}{n}\right) - f\left(\frac{1}{n}\right)\right) = f\left(\frac{2}{n}\right) - f\left(\frac{1}{n}\right).$$

Therefore, we prove these cases by examining exhaustively all possible indexes and the coordinate-wise entropy they generate.

$\underline{n=3}$**:** The possible indexes are (3), (1, 2) and (1, 1, 1), which generates the following entropy values (in bits):

$$\begin{aligned}(3) : H(Y_i) &= 0;\\ (1,2) : H(Y_i) &= f\left(\frac{1}{3}\right) = -\frac{1}{3}\log\frac{1}{3} - \frac{2}{3}\log\frac{2}{3} = \log 3 - \frac{2}{3};\\ (1,1,1) : H(Y_i) &= 3 \times \left(-\frac{1}{3}\log\frac{1}{3}\right) = \log 3.\end{aligned}$$

Once again, the only possibility for a basis to generate lower entropy than the standard basis is to include a class 1 row with index (3). But here we still cannot have two class 2 rows of index (1, 2) on top of the class 1 row since Lemma 6.4 still holds for $n = 3$. Therefore, the best combination is to have one row of each possible class, which leads to the following global coordinate-wise entropy:

$$0 + \log 3 - \frac{2}{3} + \log 3 \simeq 2.50 < 3\log 3 - 2\log 2 \simeq 2.75,$$

that is, this best possible basis is better than the standard basis. Therefore, the LSDB is a basis including a vector of each class. Considering the orthonormality of the basis, we can only have the following basis or its permuted or sign-flipped versions for $n = 3$:

$$U^T = B = \begin{bmatrix} \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{6}} & \frac{-1}{\sqrt{2}} \\ \frac{1}{\sqrt{3}} & \frac{-2}{\sqrt{6}} & 0 \end{bmatrix}.$$

$\underline{n=4}$**:** The possible indexes are: (4), (1, 3), (2, 2), (1, 1, 2), and (1, 1, 1, 1), which generate the following entropy values (in bits):

$$\begin{aligned}(4) : H(Y_i) &= 0;\\ (1,3) : H(Y_i) &= f\left(\frac{1}{4}\right) = -\frac{1}{4}\log\frac{1}{4} - \frac{3}{4}\log\frac{3}{4} = 2 - \frac{3}{4}\log 3 \simeq 0.811;\\ (2,2) : H(Y_i) &= f\left(\frac{1}{2}\right) = 1;\\ (1,1,2) : H(Y_i) &= -\frac{1}{4}\log\frac{1}{4} - \frac{1}{4}\log\frac{1}{4} - \frac{1}{2}\log\frac{1}{2} = 1.5;\\ (1,1,1,1) : H(Y_i) &= 4 \times \left(-\frac{1}{4}\log\frac{1}{4}\right) = 2.\end{aligned}$$

The total coordinate-wise entropy of the Walsh basis is $H_W \triangleq 0 + 3 \times f(1/2) = 3$ bits. We know from Theorem 5.1 that H_W is smaller than that of the standard basis. Let U be an orthonormal basis, and let $\{\boldsymbol{b}_i^T, i = 1, \ldots, 4\}$ be its rows. If U generates smaller entropy than the Walsh basis, it necessarily includes one class 1 row and one class 2 row with index (1, 3) from the same argument as the proof of Lemma 6.4 (see Appendix B). Let us assume that $\boldsymbol{b}_1^T$ is of class 1 and $\boldsymbol{b}_2^T$ of class 2 with index

$(1,3)$. In other words, $c(1) = (4)$ and $c(2) = (1,3)$. Now, $\boldsymbol{b}_2^T$ is of the form (a,a,a,b) and orthogonality with $\boldsymbol{b}_1^T$ implies that $\boldsymbol{b}_2^T$ is proportional to the vector $(1,1,1,-3)$. Now, U cannot include a class 4 row vector of index $(1,1,1,1)$. If so, these three rows (i.e., rows of class 1, 2, and 4) generate the entropy $0 + 0.811 + 2 = 2.811$ bits. Hence, any other admissible choice for the remaining row, i.e., a class 2 row with index $(1,3)$, which generates 0.811 bits, or a class 2 row with index $(2,2)$, which generates 1 bit, or a class 3 row with index $(1,1,2)$, which generates 1.5 bits, ends up larger total coordinate-wise entropy than the Walsh basis. Therefore we can discard these combinations immediately, and the indexes of $\boldsymbol{b}_3^T$ and $\boldsymbol{b}_4^T$ must be chosen from $(2,2)$ and $(1,1,2)$. If $\boldsymbol{b}_3^T$ is of index $(2,2)$, it is of the form (a,a,b,b) and orthogonality with $\boldsymbol{b}_1^T$ implies that $\boldsymbol{b}_3^T$ is proportional to $(a,a,-a,-a)$. Then, orthogonality with $\boldsymbol{b}_2^T$ implies: $a + a - a + 3a = 0$, i.e., $a = 0$. Therefore the only possibility for $\boldsymbol{b}_3^T$ and $\boldsymbol{b}_4^T$ is to be both of index $(1,1,2)$, each of which generates the coordinate-wise entropy 1.5 bits. The total coordinate-wise entropy generated by U is therefore at least $0 + 0.811 + 2 \times 1.5 = 3.811 > 3 = H_W$, hence U^T is not the LSDB. We can now conclude that the LSDB among O(4) is the Walsh Basis.

$\underline{n = 5}$**:** In this case, we prove that the LSDB is the standard basis or the basis of the Householder reflection (5.1), both of which consist of class 2 rows with index $(1,4)$ only. Indeed, using the similar argument as before, any basis generating smaller entropy than these two bases must have a class 1 row and a class 2 row with index $(1,4)$. However, since the other three rows must be either of class 2 with different indexes or of class 3 or higher, the total entropy of such a basis is larger than that of the standard basis or the Householder reflection basis:

$$\sum_{i=1}^{5} H(Y_i) \geq 0 + f\left(\frac{1}{5}\right) + 3 \times f\left(\frac{2}{5}\right) \simeq 3.635 > 5 \times f\left(\frac{1}{5}\right) \simeq 3.609.$$

This concludes the proof of Theorem 5.2. □

6.7 *Proof of Theorem 5.3*

PROOF. For the case $\mathcal{D} = \mathrm{GL}(n, \mathbf{R})$, the constraint imposed by Lemma 6.4 is lifted since the rows of $U = B^{-1}$ do not have to form an orthonormal basis anymore. Hence we can have as many rows of class 2 with index $(1, n-1)$ as we wish, even if the first row of U is of class 1. Clearly, we still cannot have two class 1 rows because this violates the invertibility of U. Therefore, considering all these remarks and the classification of indexes established in the previous subsections, it is immediate to conclude that the combination of classes of rows leading to the smallest sum of coordinate-wise entropy is one row of class 1 and $n-1$ rows of class 2 with index $(1, n-1)$. This matrix reaches the lower bound for the total coordinate-wise entropy $(n-1)f(1/n)$. Considering the invertibility

of the matrix with $n-1$ rows of class 2, the most general form of the admissible matrices is the following (modulo permutations and sign-flips mentioned in (5.2)):

$$U_{\mathrm{GL}(n,\mathbf{R})} = B^{-1}_{\mathrm{GL}(n,\mathbf{R})} = \begin{bmatrix} a & a & \cdots & \cdots & \cdots & a \\ b_2 & c_2 & b_2 & \cdots & \cdots & b_2 \\ b_3 & b_3 & c_3 & b_3 & \cdots & b_3 \\ \vdots & \vdots & & \ddots & & \vdots \\ \vdots & \vdots & & & \ddots & \vdots \\ b_{n-1} & \cdots & \cdots & b_{n-1} & c_{n-1} & b_{n-1} \\ b_n & \cdots & \cdots & \cdots & b_n & c_n \end{bmatrix},$$

where a, b_k, c_k, $k = 2, \ldots, n$, must be chosen so that $U_{\mathrm{GL}(n,\mathbf{R})} \in \mathrm{GL}(n, \mathbf{R})$. We can easily compute the determinant of this matrix in a similar manner that we derived (6.1):

$$\det\left(U_{\mathrm{GL}(n,\mathbf{R})}\right) = a \prod_{k=2}^{n} (c_k - b_k).$$

Therefore, we must have $a \neq 0$ and $b_k \neq c_k$ for $k = 2, \ldots, n$ for $U_{\mathrm{GL}(n,\mathbf{R})}$ to be in $\mathrm{GL}(n, \mathbf{R})$. Note that if we want to restrict the dictionary to $\mathrm{SL}^{\pm}(n, \mathbf{R})$, then we must have $\det\left(U_{\mathrm{SL}^{\pm}(n,\mathbf{R})}\right) = \pm 1$, i.e., a must satisfy $a = \pm \prod_{k=2}^{n} (c_k - b_k)^{-1}$.

The corresponding inverse matrix (5.3) can be computed easily by elementary linear algebra, i.e., the Gauss-Jordan method. We show this matrix here again:

$$B_{\mathrm{GL}(n,\mathbf{R})} = \begin{bmatrix} \left(1 + \sum_{k=2}^{n} b_k d_k\right)/a & -d_2 & -d_3 & \cdots & -d_n \\ -b_2 d_2/a & d_2 & 0 & \cdots & 0 \\ -b_3 d_3/a & 0 & d_3 & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0 \\ -b_n d_n/a & 0 & \cdots & 0 & d_n \end{bmatrix},$$

where $d_k = 1/(c_k - b_k)$, $k = 2, \ldots, n$. These are the LSDB pairs (analysis and synthesis respectively). This concludes the proof of Theorem 5.3. □

6.8 *Proof of Proposition 5.3*

If we transform the spike process $\boldsymbol{X}$ by the Householder reflector $B_{\mathrm{O}(n)}$ (5.1), the number of nonzero components of $\boldsymbol{Y} = B^T_{\mathrm{O}(n)} \boldsymbol{X}$ can be easily computed as

$$\mathcal{C}_0\left(B_{\mathrm{O}(n)} \,|\, \boldsymbol{X}\right) = E\|\boldsymbol{Y}\|_0 = n.$$

Now, let us consider the case $0 < p < 1$. Since $n \geq 2$, we have

$$\mathcal{C}_p\left(B_{\mathrm{O}(n)} \,|\, \boldsymbol{X}\right) = E\|\boldsymbol{Y}\|_p^p = \left(1 - \frac{2}{n}\right)^p + (n-1)\left(\frac{2}{n}\right)^p.$$

Let us now define the following function:

$$s_p(x) \triangleq (1-x)^p + \left(\frac{2}{x} - 1\right) x^p = (1-x)^p - x^p + \frac{2}{x^{1-p}},$$

where $0 < x = 2/n \leq 1$. Taking the derivative with respect to x, we have

$$s_p'(x) = -p\left(\frac{1}{(1-x)^{1-p}} + \frac{1}{x^{1-p}}\right) + \frac{2(p-1)}{x^{2-p}} < 0,$$

for $0 < x < 1$ and $0 < p \leq 1$. Therefore, in this interval, $s_p(x)$ is monotonically decreasing, and the decisive term for the sparsity measure $\mathcal{C}_p$ is $2/x^{1-p}$. Therefore, we have

$$\lim_{n\to\infty} \mathcal{C}_p\left(B_{\mathrm{O}(n)} \,|\, \boldsymbol{X}\right) = \lim_{x \downarrow 0} s_p(x) = \infty \quad \text{for } 0 < p < 1.$$

If $p = 1$, then $s_1(x) = (1-x) - x + 2 = 3 - 2x$. Hence, we have

$$\lim_{n\to\infty} \mathcal{C}_1\left(B_{\mathrm{O}(n)} \,|\, \boldsymbol{X}\right) = \lim_{x \downarrow 0} s_1(x) = 3.$$

This completes the proof. □

6.9 *Proof of Corollary 5.1*

Proof. We now consider the mutual information of the spike process under the LSDB pair (5.2) and (5.3) in Theorem 5.3, which was proved in the previous subsection. Using this analysis LSDB, the mutual information of $\boldsymbol{Y} = B^{-1}_{\mathrm{GL}(n,\mathbf{R})}\boldsymbol{X}$ is

$$\begin{aligned} I(\boldsymbol{Y}) &= -H(\boldsymbol{X}) + \sum_{i=1}^{n} H(Y_i) \\ &= -\log n + (n-1) f\left(\frac{1}{n}\right) \\ &= -\log n + (n-1)\left[\log n - \frac{n-1}{n}\log(n-1)\right] \\ &= (n-2)\log n - \frac{(n-1)^2}{n}\log(n-1). \end{aligned} \tag{6.15}$$

Let $h(n)$ denote the last expression in (6.15). Note that $h(2) = 0$, i.e., we can achieve the true independence for $n = 2$. If $n > 2$, this function is strictly positive and monotonically increasing as shown on Figures 8 and 9. By expanding the natural logarithm version of $h(x)$, we have

$$\begin{aligned} \ln 2 \times h(x) &= (x-2)\ln x - \frac{(x-1)^2}{x}\ln(x-1) \\ &= (x-2)\ln x - \left(x - 2 + \frac{1}{x}\right)\left(\ln x + \ln\left(1 - \frac{1}{x}\right)\right) \\ &= (x-2)\ln x - \left(x - 2 + \frac{1}{x}\right)\left(\ln x - \frac{1}{x} - \frac{1}{2x^2} + o\left(\frac{1}{x^2}\right)\right) \\ &= -\frac{\ln x}{x} + \left(x - 2 + \frac{1}{x}\right)\left(\frac{1}{x} + \frac{1}{2x^2} + o\left(\frac{1}{x^2}\right)\right) \\ &= 1 - \frac{\ln x}{x} - \frac{3}{2x} + o\left(\frac{1}{x}\right) \end{aligned}$$

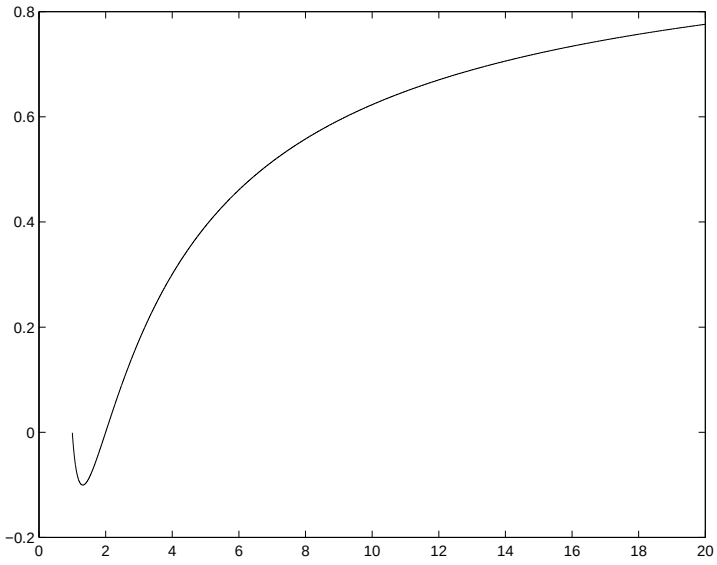

Fig. 8. A plot of the function $\ln 2 \times h : x \to (x-2)\ln x - \frac{(x-1)^2}{x}\ln(x-1)$.

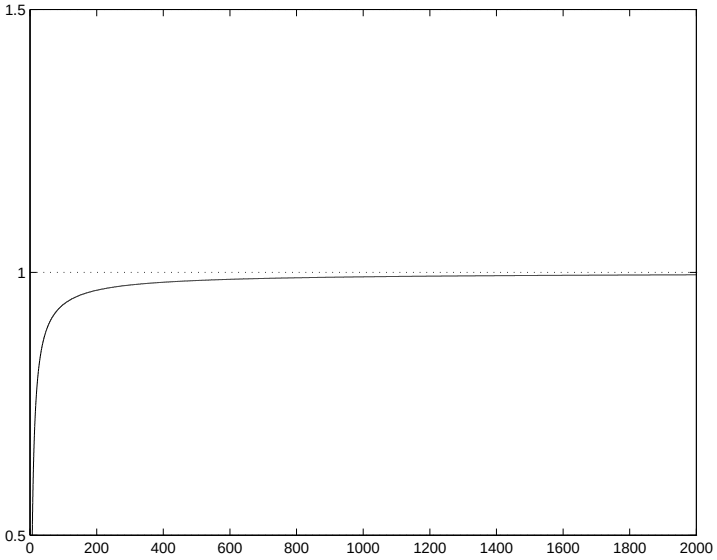

Fig. 9. A plot of the function $\ln 2 \times h(x)$ for large x.

In other words, we have established

$$I(\boldsymbol{Y}) \sim \frac{1}{\ln 2}\left(1 - \frac{\ln n}{n}\right) \quad \text{as } n \to \infty.$$

In other words,

$$\lim_{n\to\infty} I\left(B^{-1}_{\mathrm{GL}(n,\mathbf{R})}\boldsymbol{X}\right) = \frac{1}{\ln 2} = \log \mathrm{e} \approx 1.4427.$$

Therefore, for $n > 2$, there is no invertible linear transformation that gives truly independent coordinates for the spike process.

As for the orthonormal case, using (6.13), we have

$$I\left(B^T_{\mathrm{O}(n)}\boldsymbol{X}\right) = n\log n - (n-1)\log(n-1) - \log n = (n-1)\log\frac{n}{n-1} = \log\left(1 + \frac{1}{n-1}\right)^{n-1}.$$

Now, it is easy to see

$$\lim_{n\to\infty} I\left(B^T_{\mathrm{O}(n)}\boldsymbol{X}\right) = \log \mathrm{e}.$$

This completes the proof of Corollary 5.1. □

7. Discussion

In general, sparsity and statistical independence are two completely different concepts as an adaptive basis selection criterion, as demonstrated by the rotations of the 2D uniform distribution in Section 4. For the spike process, however, we showed that the BSB and the LSDB *can* coincide (i.e., the standard basis) if we restrict our basis search within $\mathrm{O}(n)$ with $n \geq 5$. However, we also showed that the standard basis is not the only LSDB in this case. To our surprise, there exists another orthonormal basis (5.1) representing the Householder reflector, which attains exactly the same level of the statistical dependence as the standard basis, if evaluated by the mutual information or equivalently by the total coordinate-wise entropy $\mathcal{C}_H$ defined in (3.3). Yet this LSDB does not sparsify the process at all *if we measure the sparsity by the expected ℓ^p norm $\mathcal{C}_p$ defined in (3.1) where* $0 \leq p \leq 1$. It is also interesting to note that this Householder reflector approaches to the standard basis as $n \to \infty$. Furthermore, if we extend our basis search to $\mathrm{GL}(n, \mathbf{R})$, then the LSDB and the BSB *cannot* coincide.

What do these observations and the effort to prove these theorems suggest? First, it is clear that proving theorems on the LSDB and computing it for more complicated stochastic processes would be much more difficult than the BSB. To deal with statistical dependency, we need to consider the "stochastics" explicitly such as entropy and the pdf of each coordinate. On the other hand, sparsity does not require such information. In fact, one can even adapt the BSB for each realization rather than for the whole realizations; see Saito et al. (2000, 2001) for further information about this issue.

Second, Remark 4 and Proposition 5.3 cast questions on the appropriateness of the ℓ^p norm $(0 \leq p < 1)$ (3.1) as a sparsity measure. According to this measure, the bases

(5.1) and (5.4) provide completely "dense" coordinates for the spike process. Yet, if we look at these basis vectors carefully, they are very "simple" in the sense that at most one component differs from all the other common components in each basis vector. In other words, *the sparsity measured by the ℓ^p norm does not imply the simplicity measured by the entropy, and vice versa.* Therefore, if a given problem really requires the statistical independence criterion, then we cannot replace it by the sparsity criterion in general.

Then, why the sparse basis of Olshausen and Field and the ICA basis of Bell and Sejnowski were more or less the same? Our interpretation to this phenomenon is the following (see also Remark 5). The Gabor-like functions they obtained essentially convert an input image patch to a spike or spike-like image. In our opinion, the image patch size such as 16×16 pixels were crucial in their experiments. Since those image patches are of small size, they tend to have simpler image contents such as simple oriented edges. It seems to us that if their algorithms were computationally feasible to accept image patches of larger size such as 64×64 or 128×128, both the BSB and the LSDB would be very different from Gabor-like functions. These large size image patches (due to rich scene variations and contents in the patches of these sizes) cannot be converted to spikes by Gabor-like simple functions.

We also note that the LSDB is not guaranteed to provide the true statistically independent coordinates in general. Therefore, if our interest is data compression, it seems to us that the pursuit of sparse representations should be encouraged rather than that of statistically independent representations. This is also the view point indicated by Donoho (1998). However, this does not mean to downgrade the importance of the statistical independence altogether. If we want to separate mixed signals or to build empirical models of stochastic processes for simulation purposes, then pursuing the statistical independence should be encouraged, and we expect to see further interplay between these two criteria.

Finally, there are a few interesting generalizations of the spike process, which need to be addressed in the future. One is the spike process with varying amplitude. The spike process whose amplitude obeys the normal distribution was treated by Donoho et al. (1998) to demonstrate the superiority of the non-Gaussian coding using spike location information over the Gaussian-KLB coding. The other generalization is to randomly throw in multiple spikes to a single realization. If one throws in more and more spikes to one realization, the standard basis is getting worse in terms of sparsity. It will be an interesting exercise to consider the BSB and the LSDB for such situations.

References

Bell, A. J. and Sejnowski, T. J. (1997). The 'independent components' of natural scenes are edge filters, *Vision Research*, **37**, 3327–3338.

Cardoso, J.-F. (1999). High-order contrasts for independence component analysis, *Neural Computation*, **11**, 157–192.

Coifman, R. R. and Wickerhauser, M. V. (1992) Entropy-based algorithms for best basis selection, *IEEE Trans. Inform. Theory*, **38**, 713–719.

Cover, T. M. and Thomas, J. A. (1991) *Elements of Information Theory*, Wiley Interscience, New York.

Daubechies, I. (1992) *Ten Lectures on Wavelets*, SIAM, Philadelphia, PA.

Day, M. M. (1940). The spaces L^p with $0 < p < 1$, *Bull. Amer. Math. Soc.*, **46**, 816–823.

Donoho, D. L. (1994). On minimum entropy segmentation, Technical Report, Dept. Statistics, Stanford University.

Donoho, D. L. (1998). Sparse components of images and optimal atomic decompositions, Technical report, Dept. Statistics, Stanford University.

Donoho, D. L., Vetterli, M., DeVore, R. A., and Daubechies, I. (1998). Data compression and harmonic analysis, *IEEE Trans. Inform. Theory*, **44**, 2435–2476.

Hall, P. and Morton, S. C. (1993). On the estimation of entropy, *Ann. Inst. Statist. Math.*, **45**, 69–88.

Lin, J.-J., Saito, N., and Levine, R. A. (2000) An iterative nonlinear Gaussianization algorithm for resampling dependent components, *Proc. 2nd International Workshop on Independent Component Analysis and Blind Signal Separation*, 245–250, IEEE.

Lin, J.-J., Saito, N., and Levine, R. A. (2001) An iterative nonlinear Gaussianization algorithm for image simulation and synthesis, preprint.

Olshausen, B. A. and Field, D. J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images, *Nature*, **381**, 607–609.

Olshausen, B. A. and Field, D. J. (1997). Sparse coding with an overcomplete basis set: A strategy employed by V1? *Vision Research*, **37**, 3311–3325.

Saito, N. (2000). Local feature extraction and its applications using a library of bases, *Topics in Analysis and Its Applications: Selected Theses* (ed. R. Coifman), 269–451, World Scientific, Singapore.

Saito, N. (2001). Image approximation and modeling via least statistically-dependent bases, *Pattern Recognition*, to appear.

Saito, N., Larson, B. M., and Bénichou, B., (2000). Sparsity and statistical independence from a best-basis viewpoint, *Proc. SPIE*, **4119**, *Wavelet Applications in Signal and Image Processing VIII* (eds. A. Aldroubi, A. F. Laine, and M. A. Unser), 474–486, Invited paper.

Saito, N., Bénichou, B., and Larson, B. M. (2001). Sparsity and statistical independence in adaptive signal representations. In preparation.

van Hateren, J. H. and van der Schaaf, A (1998). Independent component filters of natural images compared with simple cells in primary visual cortex, *Proc. Royal Soc. London, Ser. B*, **265**, 359–366.

Wickerhauser, M. V. (1994). *Adapted Wavelet Analysis from Theory to Software.* A K Peters, Wellesley, MA.

Appendix A: Proof of Lemma 6.2

PROOF. First we need to show another lemma as follows:

LEMMA A.1. *Let $p_2 \geq p_1 \geq 1$ be positive integers such that $p_1 + p_2 \leq n$. Then*

$$\frac{p_1}{n}\log\frac{p_1}{n} + \frac{p_2}{n}\log\frac{p_2}{n} \leq \frac{p_1+p_2}{n}\log\frac{p_1+p_2}{n} - \frac{2}{n}f\left(\frac{1}{n}\right),$$

where f is defined in (6.6).

PROOF. The left-hand side of the inequality can be written as

$$\begin{aligned}
\frac{p_1}{n}\log\frac{p_1}{n} + \frac{p_2}{n}\log\frac{p_2}{n} &= \left(\frac{p_1+p_2}{n}\right)\left[\frac{p_1}{p_1+p_2}\log\frac{p_1}{n} + \frac{p_2}{p_1+p_2}\log\frac{p_2}{n}\right] \\
&= \left(\frac{p_1+p_2}{n}\right)\left[\log\frac{p_1+p_2}{n} + \frac{p_1}{p_1+p_2}\log\frac{p_1}{p_1+p_2} + \frac{p_2}{p_1+p_2}\log\frac{p_2}{p_1+p_2}\right] \\
&= \left(\frac{p_1+p_2}{n}\right)\log\left(\frac{p_1+p_2}{n}\right) + \left(\frac{p_1+p_2}{n}\right)\left[-f\left(\frac{p_1}{p_1+p_2}\right)\right] \qquad \text{(A.1)}
\end{aligned}$$

However, it is clear that

$$\frac{1}{2} \geq \frac{p_1}{p_1+p_2} \geq \frac{1}{p_1+p_2} \geq \frac{1}{n}.$$

From the monotonicity of $f(x)$ for $x \in [0, 1/2]$, we deduce

$$1 = f\left(\frac{1}{2}\right) \geq f\left(\frac{p_1}{p_1+p_2}\right) \geq f\left(\frac{1}{n}\right),$$

which we can rewrite as

$$-1 \leq -f\left(\frac{p_1}{p_1+p_2}\right) \leq -f\left(\frac{1}{n}\right).$$

This inequality, nonnegativity of f, and the assumption of this lemma yields

$$\left(\frac{p_1+p_2}{n}\right)\left[-f\left(\frac{p_1}{p_1+p_2}\right)\right] \leq -\frac{2}{n}f\left(\frac{1}{n}\right).$$

This inequality combined with (A.1) completes the proof of Lemma A.1. □

Coming back to the proof of Lemma 6.2, we now use induction as follows.

<u>$k = 3$:</u> Since $\alpha_1 + \alpha_2 < n$, we can use Lemma A.1 to assert

$$\frac{\alpha_1}{n}\log\frac{\alpha_1}{n} + \frac{\alpha_2}{n}\log\frac{\alpha_2}{n} \leq \frac{\alpha_1+\alpha_2}{n}\log\frac{\alpha_1+\alpha_2}{n} - \frac{2}{n}f\left(\frac{1}{n}\right).$$

Therefore,

$$\begin{aligned}
\sum_{j=1}^{3}\frac{\alpha_j}{n}\log\frac{\alpha_j}{n} &\leq \frac{\alpha_3}{n}\log\frac{\alpha_3}{n} + \frac{\alpha_1+\alpha_2}{n}\log\frac{\alpha_1+\alpha_2}{n} - \frac{2}{n}f\left(\frac{1}{n}\right) \\
&= \frac{\alpha_3}{n}\log\frac{\alpha_3}{n} + \left(1-\frac{\alpha_3}{n}\right)\log\left(1-\frac{\alpha_3}{n}\right) - \frac{2}{n}f\left(\frac{1}{n}\right) \\
&= -f\left(\frac{\alpha_3}{n}\right) - \frac{2}{n}f\left(\frac{2}{n}\right).
\end{aligned}$$

We used the fact $\sum_{j=1}^{3}\alpha_j = n$ to derive the equality in the second line of the above expression. Since $\alpha_j \geq 1$ for $j = 1, 2, 3$, we must have $(n-1)/n > \alpha_3/n \geq 1/n$. Considering the symmetry of $f(x)$ around $x = 1/2$ and its behavior, we can deduce that

$$\sum_{j=1}^{3}\frac{\alpha_j}{n}\log\frac{\alpha_j}{n} \leq -f\left(\frac{1}{n}\right) - \frac{2}{n}f\left(\frac{2}{n}\right) \leq -\left(1+\frac{2}{n}\right)f\left(\frac{1}{n}\right).$$

This nails down the case $k = 3$.

$\underline{k \Rightarrow k+1}$**:** Let us demonstrate that, assuming that the formula is true for $k \geq 3$, it is still true for $k+1$. We can decompose the sum $\sum_{j=1}^{k+1} \frac{\alpha_j}{n} \log \frac{\alpha_j}{n}$ in the following way:

$$\sum_{j=1}^{k+1} \frac{\alpha_j}{n} \log \frac{\alpha_j}{n} = \frac{\alpha_{k+1}}{n} \log \frac{\alpha_{k+1}}{n} + \frac{\alpha_k}{n} \log \frac{\alpha_k}{n} + \sum_{j=1}^{k-1} \frac{\alpha_j}{n} \log \frac{\alpha_j}{n}.$$

But once again, since $\alpha_k + \alpha_{k+1} < n$, we can use Lemma A.1 to reach

$$\frac{\alpha_{k+1}}{n} \log \frac{\alpha_{k+1}}{n} + \frac{\alpha_k}{n} \log \frac{\alpha_k}{n} \leq \frac{\alpha_{k+1} + \alpha_k}{n} \log \frac{\alpha_{k+1} + \alpha_k}{n} - \frac{2}{n} f\left(\frac{1}{n}\right). \tag{A.2}$$

Let us rename a sequence $\{\alpha_j\}$ as follows:

$$\beta_j = \begin{cases} \alpha_{j+1} + \alpha_j & \text{if } j = k; \\ \alpha_j & \text{if } j = 1, \ldots, k-1. \end{cases}$$

Then, using the induction assumption, we can rewrite (A.2) as

$$\sum_{j=1}^{k+1} \frac{\alpha_j}{n} \log \frac{\alpha_j}{n} \leq \frac{\beta_k}{n} \log \frac{\beta_k}{n} + \sum_{j=1}^{k-1} \frac{\beta_j}{n} \log \frac{\beta_j}{n} - \frac{2}{n} f\left(\frac{1}{n}\right).$$

Since $\sum_{j=1}^{k} \beta_j = \sum_{j=1}^{k+1} \alpha_j = n$, we can state that

$$\begin{aligned} \sum_{j=1}^{k+1} \frac{\alpha_j}{n} \log \frac{\alpha_j}{n} &\leq \sum_{j=1}^{k} \frac{\beta_j}{n} \log \frac{\beta_j}{n} - \frac{2}{n} f\left(\frac{1}{n}\right) \\ &\leq -\left(1 + \frac{2(k-2)}{n}\right) f\left(\frac{1}{n}\right) - \frac{2}{n} f\left(\frac{1}{n}\right) \\ &= -\left(1 + \frac{2(k-1)}{n}\right) f\left(\frac{1}{n}\right). \end{aligned}$$

This concludes the proof of Lemma 6.2. □

Appendix B: Proof of Lemma 6.4

PROOF. Let us prove this lemma with *reductio ad absurdum*. Let us assume that, for example, $c(2) = c(3) = (1, n-1)$. Since the first row of U is proportional to $(1, 1, \ldots, 1)$, all the other rows must satisfy $\sum_{j=1}^{n} u_{ij} = 0$ for $i = 2, \ldots, n$ because of the orthonormality condition. Let us now consider the second row $(u_{21}, \ldots, u_{2n})$. Since $c(2) = (1, n-1)$, let us assume $u_{21} = a$ and $u_{2j} = b$, $j = 2, \ldots, n$ for some $a, b \in \mathbf{R}$. Then the orthonormality condition implies $a + (n-1)b = 0$. Since the norm of this row vector has to be one, we also have $a^2 + (n-1)b^2 = 1$. From these two constraints, we have $(n-1)^2 b^2 + (n-1)b^2 = 1$. This implies $a = \pm\sqrt{\frac{n-1}{n}}$ and $b = \mp\frac{1}{\sqrt{n(n-1)}}$.
As the second and third rows of U must be linearly independent, we need to assume that the third row is $(c, d, c, \ldots, c)$ for some $c, d \in \mathbf{R}$. (We cannot assume $(d, c, \ldots, c)$ for the third row since its inner product with the second row gives $ad + (n-1)bc = 0$, which leads to $c = d$ using the values of a and b obtained above.) Then, similarly to the second row, we also get $d + (n-1)c = 0$, $d^2 + (n-1)c^2 = 1$. Thus, we have $d = \pm a$ and $c = \pm b$. Then, regardless of the choice of the signs for a, b, c, d, the orthogonality of the second and third rows yields

$$0 = (n-2)b^2 + 2ab = (n-2) \cdot \frac{1}{n(n-1)} - 2 \cdot \frac{1}{n}.$$

This leads to $2 = \frac{n-2}{n-1}$, i.e., $2n - 2 = n - 2$, and finally $n = 0$. This contradiction implies that the assumption made is impossible, and proves the lemma. □

Appendix C: Proof of Lemma 6.6

PROOF. Our strategy of proving this lemma is the following. First we will show that the LSDB selected from $\mathrm{O}(n)$, which consists of only class 2 row vectors with index $(1, n-1)$, must be of the form:

$$\begin{bmatrix} a_1 & b_1 & \cdots & \cdots & \cdots & b_1 \\ b_2 & a_2 & b_2 & \cdots & \cdots & b_2 \\ \vdots & & \ddots & & & \vdots \\ \vdots & & & \ddots & & \vdots \\ b_{n-1} & \cdots & \cdots & b_{n-1} & a_{n-1} & b_{n-1} \\ b_n & \cdots & \cdots & \cdots & b_n & a_n \end{bmatrix}. \tag{C.1}$$

where $a_k^2+(n-1)b_k^2 = 1$ for $k = 1, \ldots, n$. We then derive the final form (5.1) using the orthonormality of the row vectors of this matrix (C.1).

Since each row is of class 2 with index $(1, n-1)$, only one the entry in a row must be different from all the other $n-1$ entries. Therefore, without loss of generality, in the kth row, let a_k be such a distinguishing entry and b_k be the other $n-1$ entries. Let $B = U^T$ be the LSDB under consideration. Suppose U has the ith and jth rows in which the locations of a_i and a_j coincide. Without loss of generality (modulo row and column permutations), we can assume that U is of the following form.

$$\begin{bmatrix} a_1 & b_1 & \cdots & \cdots & \cdots & b_1 \\ a_2 & b_2 & \cdots & \cdots & \cdots & b_2 \\ b_3 & a_3 & b_3 & \cdots & \cdots & b_3 \\ \vdots & & \ddots & & & \vdots \\ b_{n-1} & \cdots & \cdots & a_{n-1} & b_{n-1} & b_{n-1} \\ b_n & \cdots & \cdots & b_n & a_n & b_n \end{bmatrix}. \tag{C.2}$$

From the normalization condition, we must have:

$$a_k^2 + (n-1)b_k^2 = 1 \quad \text{for } k = 1, \ldots, n. \tag{C.3}$$

From the orthonormality condition, $U^TU = I_n$, the diagonal entries of U^TU are:

$$\begin{aligned} (U^TU)_{1,1} &= 1 = a_1^2 + a_2^2 + \sum_{j=3}^{n} b_j^2, \\ (U^TU)_{k,k} &= 1 = a_k^2 + \sum_{j=1, j\neq k}^{n} b_j^2, \quad 2 \le k < n, \\ (U^TU)_{n,n} &= 1 = \sum_{j=1}^{n} b_j^2. \end{aligned}$$

These imply that $a_k^2 = b_k^2$ for $k \ge 3$. Inserting this to (C.3) and noting that we must have $a_k \neq b_k$ because of the class 2 condition, we obtain:

$$a_k = \pm 1/\sqrt{n},\ b_k = \mp 1/\sqrt{n}, \quad \text{for } k \ge 3. \tag{C.4}$$

Consider now the off-diagonal entry of U^TU, for example,

$$\begin{aligned} (U^TU)_{1,2} &= 0 = a_1b_1 + a_2b_2 + a_3b_3 + b_4^2 + \cdots + b_n^2, \\ (U^TU)_{1,n} &= 0 = a_1b_1 + a_2b_2 + b_3^2 + b_4^2 + \cdots + b_n^2 \end{aligned}$$

Inserting (C.4) into these, we get

$$a_1 b_1 + a_2 b_2 - \frac{1}{n} + \frac{n-3}{n} = 0$$
$$a_1 b_1 + a_2 b_2 + \frac{n-2}{n} = 0.$$

This is a contradiction (i.e., $a_1 b_1 + a_2 b_2$ cannot have two different values). Therefore U cannot have two rows where the distinguishing entries a_i, a_j share the same column index as (C.2). It is clear that we cannot have more than two such rows. Therefore, U must be of the form (C.1).
Now, let us compute the entries of (C.1). The normalization condition (C.3) still holds. Computing the diagonal entries of $U^T U = I_n$, we have

$$(U^T U)_{k,k} = 1 = a_k^2 + \sum_{j=1, j\neq k}^{n} b_j^2 \quad \text{for } k = 1, \ldots, n. \tag{C.5}$$

Combining (C.3) and (C.5), we have:

$$n b_k^2 = \sum_{j=1}^{n} b_j^2 \quad \text{for } k = 1, \ldots, n.$$

This implies that $b_1^2 = \cdots = b_n^2$. Then, from the normalization condition (C.3), we must have $a_1^2 = \cdots = a_n^2$ also. Consider now the off-diagonal entry of $U^T U$:

$$(U^T U)_{1,2} = 0 = a_1 b_1 + a_2 b_2 + (n-2) b_1^2.$$

Now, we must have $b_2 = b_1$ or $b_2 = -b_1$. So, the above equation can be written as

$$(U^T U)_{1,2} = 0 = a_1 b_1 \pm a_2 b_1 + (n-2) b_1^2.$$

This implies that either $b_1 = 0$ or $a_1 \pm a_2 + (n-2) b_1 = 0$. $b_1 = 0$ leads to $b_k = 0$ and $a_k = \pm 1$, i.e., the standard basis. Let us consider now the other case, i.e., $a_1 \pm a_2 + (n-2) b_1 = 0$. Since $a_2 = a_1$ or $a_2 = -a_1$, these lead to either $b_1 = 0$ or $2a_1 + (n-2) b_1 = 0$. The former case has been already treated. Thus, let us proceed the latter case. From this, we have

$$a_1 = \left(1 - \frac{n}{2}\right) b_1. \tag{C.6}$$

Inserting this into (C.3), we have

$$b_1^2 = \frac{4}{n^2}.$$

Consequently,

$$a_1^2 = 1 - (n-1) \cdot \frac{4}{n^2} = \left(\frac{n-2}{n}\right)^2.$$

Because of (C.6) (that is true for all k), we have:

$$a_k = \pm \frac{n-2}{n},\ b_k = \mp \frac{2}{n}, \quad \text{for } k = 1, \ldots, n. \tag{C.7}$$

This means that the matrix U must be of the following form or its permuted and sign-flipped versions:

$$U = \frac{1}{n} \begin{bmatrix} n-2 & -2 & \cdots & -2 \\ -2 & n-2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & -2 \\ -2 & \cdots & -2 & n-2 \end{bmatrix}.$$

It turns out that this is symmetric, so we have $B = U$. This completes the proof of Lemma 6.6. □